Interviewer: So, I will start by telling you a bit about this project and also about how the interview will be conducted. Thank you so much for agreeing to participate. It's really valuable.

Participant: No problem.

Interviewer: The aim of the project is to map how researchers work with research data today and what their needs are regarding research data. The goal is to develop the best possible support for researchers so that you can focus on your research. And the discussion today will be structured around the simplified model of working with research data. And I will show you what it looks like (…) Can you see the image?

Participant: I can see the image, yes.

Interviewer: This model consists of different phases: planning, data collection, analysis, sharing, preservation, and reporting. And we will look specifically at something called capabilities. It has nothing to do with a person's capabilities, but rather it's a broad concept that describes processes, information resources, competences, regulations, IT support services, and so on in relation to the e-infrastructures used by research. So, that is all the support services that are needed to be able to manage research data well through all phases of a research activity. We will not look at all processes connected to research today but focus on what e-infrastructure is needed to manage research data well.

Participant: Okay.

Interviewer: And so, this is what we mean by e-infrastructure [shows image]. You don't need to memorize this picture. It's just an illustration to show some examples of things that could be considered e-infrastructure. At the lowest level, it can be the systems and networks that you need to store and work with data. It can also be the infrastructures and tools that you use to collect or analyze data. It can be platforms that you use to collaborate around data or to publish data. And at the very top are some examples of support services that you might use or have access to. And all of this is the e-infrastructure. It might not be something that you think about in your daily work. But you

think about it when it's not available or it doesn't work [chuckles], because it's necessary for your work as a researcher. Please keep this image in mind during the interview, as it represents the environment that you as a researcher need to relate to and move in during your everyday work.

Participant: Yep, okay.

Interviewer: The interview goes like this: we will start from a research activity that is relevant for you, perhaps a currently recently completed project because it's usually easier to talk about a specific case, but if you come up with comments of a more general kind these are also very welcome. I will stop sharing now and [notetaker] will show you an image.

Notetaker: So, there you go.

Interviewer: This is the same wheel with the different phases that I showed you earlier but now it has space for writing so you can tell us about how you work with data in the different stages but you don't have to stick to them strictly you can speak freely. [Notetaker] will briefly note what you tell us on the picture of with the wheel and we will check that, in the end we will check if you think that we have caught what you have said correctly and you can correct if you need to or add anything. [chuckles]

Participant: Alright.

Interviewer: In the end we will ask you to indicate which of the figures up in the left-hand corner that you would choose to represent your overall satisfaction for each of the phases: happy for good, neutral for well or frowning for dissatisfied and it's the overall impression of how it works to work with data in the different phases. So, it's not about if you're happy with your research but rather are you happy with the digital environment and the digital tools.

Participant: Can I just ask a quick question?

Interviewer: Yes, absolutely.

Participant: Because I guess you're asking researchers that are from all different departments, so there I would reckon that the digital environment is going to be varying across departments too, right?

Interviewer: Yes, yeah, we're asking researchers at different departments and also different universities and different research disciplines.

Participant: Okay.

Interviewer: It's going to be, yes, it's probably going to be very different.

Participant: Yeah, it's going to be quite variable, I would assume. Yeah, okay.

Interviewer: As I mentioned earlier, we will record the audio during the interview and it will be transcribed, after which the recording will be erased.

Participant: Okay.

Interviewer: What is analyzed is the transcription of the interview, the text, and the notes that we take during the interview.

Participant: I will only curse for emphasis occasionally then!

[both laugh]

Interviewer: And [Notetaker] you can choose if you note that down or not.

[all laugh]

Notetaker: Yeah.

Interviewer: So, do you have any more questions before we get started?

Participant: I think I get the general idea. I think maybe once you guys start asking maybe more specific questions, I'll be able to kind of get into the groove a little bit, but I think I get the idea. Yeah.

Interviewer: Yeah. Okay, good. So first, could you briefly tell us about your research in general with focus on the data that you collect or generate?

Participant: So, working in the field of microbial ecology, and I mean, myself, I primarily work with soil microbial communities. And to be a bit more specific, I'm working quite a lot with organisms that are involved in biogeochemical cycling, mainly the nitrogen cycle. What(--) I guess to relate this to the data, when we're doing any sort of study, we're collecting essentially a range of different environmental data, and we try to get some level of activity data for different assays that we perform to understand how the microbes are responding to different treatments in the field or different treatments in the lab and different kinds of experiments. And we also collect a lot of sequence data in order to do profiling of microbial communities. And this can be sequence data that's related to sort of standard taxonomic profiling. But then also we look at other types of marker genes or different groups of microbes. And let's see, [hesitates] but that is relating to, let's say, field or lab experiments.

We also do a lot of sort of just pure data-driven work where we can go into large public databases like NCBI or ENA, EMBL, JGI, and download a lot of sequence data from, for example, metagenomes or genomes, and then do comparative work in which we will, yeah, basically try to collate a large database, sequence database, along with associated metadata and try to derive, I guess you could say interesting patterns from the data that's just been sort of shoved up into these public repositories and it becomes kind of a, you know, the usual cycle of you have hypothesis driven research and you have hypothesis generating research. We cycle back and forth between those two types of modes of working I guess you could say. I think that about sums it up.

Interviewer: Yeah, so if you would pick a specific project. Do you have one in mind?

Participant: A specific project?

Interviewer: Something you have done recently or something that you–

Participant: [interrupts] That we completed recently. We [hesitates] I was involved in a couple of studies where we were doing exactly what I just described when it comes to downloading, it was about 3,000 metagenomes from different environments that were obtained from different environments, different biomes, if you will. And we were interested in looking at essentially the co-association of specific genes that are involved

in, in this case, it's a nitrogen cycle process called denitrification, which is converting nitrate into either nitrous oxide or dinitrogen gas and we're of course interested in looking at whether or not you wind up seeing different patterns of whether the abundances of these genes tend to be more correlated across different environments under different environmental conditions–

Interviewer: [interrupts] So, these were data that you collected from a database, these were not

Participant: Yes.

Interviewer: collected?

Participant: Yes.

Interviewer: From the field, yes okay.

Participant: Yeah, we collected this from NCBI primarily and also JGI. We spent God knows how long downloading it because that's always the biggest hurdle, getting the data. Because yeah, you know I mean if you're trying to ping the server over in Washington for NIH NCBI it, you know, it shuts off from time to time and the connection can be really lousy. But we're also like going through that data and manually curating it because there's just a lot of garbage out there so we have to sift through it and it's tedious but that's just part of how you make a good database that you can do the subsequent work on where you want to detect and find the genes, make sure that they're correctly annotated, the ones you find, and then you can start doing the actual science part where you're basically doing statistical analysis to see what particular patterns are robust, which ones are holding, what particular regression models might be, you know, what the more important variables are for predicting certain patterns in a model.

Interviewer: So, when you download these datasets do you store them on your own computer locally or do you have a database at a server or where do you store the data that you have downloaded?

Participant: In this department we have a computing grid, we have a server that has been basically developed by other researchers within the department. We've had it for, I mean, I don't know, God knows how many years. Our group has been in this department for, I think, about 11 years, but they had this beforehand. But it's basically a large computing grid that we've been able to use. It's very well constructed, it's very well maintained, and it's just for our department, so it's kind of a luxury. And we have, I don't know how many hundreds of terabytes of storage space that we can then download and just simply maintain a large metagenomic database from all these different studies that we're collecting. It(--) we do hog up a lot of space [laughs] on the grid, to our IT coordinators chagrin but at the same time they're very good at trying to keep up with the storage needs and the computing needs. So yeah, that's basically(--) in our department I think we have it fairly good, I would say.

I can also maybe add that we have also been utilizing some of NAISS's resources. So the computing infrastructure, we used to be in UPPMAX before they got rid of it. That was unfortunate when that happened so we are occasionally using Dardel, but Dardel and [university] isn't exactly the nicest machine to use, so we've often just simply come back to our own grid and we like having that kind of just as a within house resource. I would say we're 90 percent, 95 percent using our in-house resources, our in-house grid.

Interviewer: And is that, it's not connected to [university]'s IT?

Participant: No it's, well I mean, it's run by the department IT but it's not it's not connected to like the head IT, I mean I'm sure there's some sort of you know connection just because of the security and everything like that and I know that the IT coordinator in our department is usually having regular meetings with the head IT people but it's more just to try to(--) I shouldn't speak for [them] I'm not exactly sure, I hear [them] complain quite a bit [chuckles] you can leave it at that.

Interviewer: Do you know if the data is backed up somehow?

Participant: Yes. Yeah we have an in-house backup system so that way the data that we download, there is another storage server in which the entirety of the grid(--)

because it's not just our projects there's other projects for other groups within the department and they have the grid on a dual, on another backup system as well.

Interviewer: Okay.

Participant: We do have, because of the size of our database and the dataset, we do have a bit of an issue when it comes to some of the analyses, where we have to sort of be a bit more judicial on what we want backed up. Because when you do these types of analyses you can get a large number of, we'll call them intermediate files, they're files are being produced during the analysis that, they're not essential to keep, it's not like it's the main data source and they can get quite large. So we try to keep those but otherwise, yeah like for the main data that we download and the source data that's backed up.

Interviewer: Okay so in this part of the work have you used any support functions for example [support function at university] or SciLifeLab or have you?

Participant: No, we pretty much been keeping it in-house [hesitates] I think we have tried to use [support function at university] once a long time ago, but I think what we were(--) it didn't, it didn't work out and through no fault of the person that we hired or consulted with. What we were doing was rather specialized, so we had to sort of develop our own way of doing it through just basically figuring out with different scripts and various sanity checks to make sure that what we had was what we were looking for and so on.

Notetaker: May I ask a question, [interviewer]?

Participant: Sure.

Interviewer: Yeah, of course.

Notetaker: Have you used, I can't remember now the official name, but these, what are they called? At SciLifeLab, they're part of Wallenberg's.

Participant: Yeah, yeah, yeah.

Interviewer: Kind of the support functions there?

Participant: Yeah, NB, NIBIO, N [hesitates] I always forget the name of it too, but we actually have.

Notetaker: NBIS support.

Participant: NBIS, NBIS, NBIS. Yeah. We have planned to, we haven't yet. But a former colleague from within the group is working there now. And we, for a project that we're working on now, we have planned to basically hire like a long-term support. And that's primarily because we're starting to dabble more with, I hate calling it AI because AI is the wrong term. Computational statistics. Machine learning and different types of approaches. In order to make sure that we basically have the right idea and that it can be implemented, we plan to send in an application for NBIS support. But we haven't yet.

Interviewer: Is that for the analysis of the data or also for the planning?

Participant: Yes, that'll be for the analysis of the data. The current planning basically follows along with the previous project that I described we're just expanding the database because I mean data is being uploaded you know exponentially, and the studies that I use as an example there they were started back in 2021-22 and essentially one of the last papers I mean. But we're still updating it and constantly using it because we can always pull out something new and something interesting.

Interviewer: The plan, how did you work with the planning of the project? You said it builds on previous work.

Participant: Yeah.

Interviewer: Did you(--) How did you plan the data management?

Participant: How do we plan the data management? Basically what we did is, I mean it's kind of like the standard way that you do a meta study, is you're going through a lot of different papers but then also going through databases where you have specific search terms for let's say, now let me back up [hesitates] we had to set specific criterion so that the data was kind of normalized, so that we could make comparable, we could have comparable datasets across all the different studies that we were collecting. For

example, that means if, because there's a lot of different sequencing technologies for example, so we had to pick one. That it would be sort of like, okay, we stick with paired and Illumina sequencing, there's a lot of other sequencing options that were available, but we wanted to at least say okay but there might be a bias if you have one platform versus another and we wanted to keep the question simple so we set that criterion and then we had other criterion where we could get enough metadata where we could actually describe the biome type in a standardized form.

And then when it comes to managing the data once you have these accession numbers that you can get for different metagenomes we were making sure that we had essentially a SQL database, so you know SQL set up where we have multiple tables and then we could get the metadata and then associate that with(--) Because the sequence files are huge right, they can be very, very large there's like you know tens of hundreds of gigabytes sometimes so, well more like tens really. But essentially, we would just make sure that we could always trace back an accession number to a given set of metadata and we would(--) we had a SQL database and we're further developing that now but that can also, whenever, for example, when we publish using this dataset, it's easy to provide all the information in an Excel file as well. But the SQL just provides you with a bit more flexibility on data to use, for example.

Interviewer: So, in this project, you say "we", so are you collaborating with someone, within the department, or outside of your in another organization?

Participant: No, it's within the group. So "we" being, at the time, a couple postdocs, a graduate student, and of course the group leader, the PI, is heavily involved. And basically, it's just been within the group collaborative effort to work with this data. We had on one paper, or was it two, we did have other collaborators listed for example in the author list, but it was more for contributing to just sort of the scientific ideas, not necessarily the data collection in and of itself.

Interviewer: So how is this project funded?

Participant: Oh boy, I just I can't remember right now I use your typical [funders] grants. I believe that this one was [funder]. But it was a long-standing effort because it did go

back even further than 2021, because the idea for it was sort of drummed up maybe a couple years before that. And because it was so difficult, it kind of fermented for a little bit. It had to age a little bit. And then once we started, basically when we hired additional postdocs to really help and start driving it more, it sort of got off the ground and then now it's been published. At least a couple of the studies have been published. But like I said, this is still, this is a resource that it's not just a one-time thing, it's something that we're continuously using.

Interviewer: When you applied for funding, did you also apply for money for that relates to the data management or data collection specifically?

Participant: I was not the one that applied for it. The PI was, but I don't recall that being part of the grant application as far as budgeting goes. But I can't give you an absolute answer on that, but I don't think so. I think because basically the way it works when it comes to data management in this department, when you apply for any grant, you have to put an overhead, right? So the overhead, that's kind of what you're paying for with this department is the technical support.

Interviewer: Yeah, okay.

Participant: Probably the best way to put it that I can think of.

Interviewer: So did you encounter any problems or difficulties in planning the project. The planning phase?

Participant: Besides the usual?

[both laugh]

Interviewer: Yeah, yeah.

Participant: Yes, it's a science, yes, it's a science project it was a royal pain in the ass [laughs], but I mean it's more about(--) and then that's just the way bioinformatics works, there's always going to be issues with some package software script that you wrote and then you just got to problem solve until you get it. When it comes to difficulties with the infrastructure per se in the department or [university] really it was just more about(--) I

mean the at the time I remember we had a hell of a time, because the bandwidth for the internet was pretty low. So we would have, it'll be very slow downloading stuff and we would have very slow speeds. And I mean this is something that I think in [university] like everyone's kind of having a bit, because they need to upgrade their network and I think(--) As I understand this is a more anecdotal second hand. But yeah, it's been the network has been problematic at [university]. We kind of, we were affected by say, because it was quite tedious to have to work with downloading a lot of this data.

Interviewer: And you also said you had some issues with the quality of the data in the databases that you have.

Participant: Well, I mean, yeah, but that's nothing to do with infrastructure, that's just the nature of data that's being uploaded into these databases. Because it's always [hesitates] you always have to be, you know keep a critical eye and be skeptical about what it is that you're seeing and where it came from so, that's just kind of, that's life.

Interviewer: Yeah.

Participant: So, you just have to be really careful about what you want to keep in your database and what should be thrown away. That's just that's just common everywhere.

Interviewer: Yeah.

Notetaker: I was wondering something you said you have(--) First of all you have got quite a good kind of departmental IT support and I guess you have very good communication between them. What(--) But when you were planning for this did you include them, was there kind of a discussion "oh well, right now we just have this amount of storage open, or"?

Participant: When we were planning, yeah, it wasn't like we sat down and had a giant formal meeting. Because it's, when we were planning we simply planned around, okay let's say that we want to actually design this particular project and then we say we want to try to download this and download that. Then we would plan that out like with the student, because we had another student project that was involved with another paper using the same database. And then to see what we could do we would go and meet up

with the IT people and say how much storage can we get, what is going to be feasible and it would be just, like, working in the lab you're problem solving back and forth between IT people and the people that are in charge of the computing grid.

And to be honest what happened was before we started this big project there was only one person that was actually, two people rather, that were allowed to install software on the grid for example. But our group, because we have a little bit more competence with computing, those two individuals that were administrating the grid they allowed people in our group to start installing software because we knew how to do it. So it, I would say in the end there was definitely a communication back and forth. It just didn't, it wasn't like "okay, let's sit down and plan" it was like "okay, what can we do, how can we build stuff up and then we will go back and forth over time to continuously build up the competence and build up the system too". Because they would also say like "oh, yeah, but we're going to be buying this much storage maybe we should get more" like get more, that kind of thing.

Interviewer: What kind of software were you using for this project?

Participant: I mean it's all bioinformatics software so it could be anything that's working with like hidden Markov models or doing sequence searches. Because we're primarily working with marker genes and the way we do it (.) is we are trying to [hesitates] what's the best way to explain it so it's not super nerdy and technical (.) the hardest part with looking for genes from environmental samples is trying to make sure that the hits you get, what you find, is the actual protein that you're interested in because there could be other ones that are closely related but they're not doing what it is, they're you [hesitates] they're not doing what it is you're interested in.

For software we'll be using, you know, standard sequence searching software this could be like BLAST, like DIAMOND, like HMMER, hidden Markov models but then also a lot of phylogeny. So, phylogenetic analysis where you're trying to essentially make trees, showing how related sequences are to each other. And then from that, you can, I detect, okay, you can use a combination of your sequence search and a phylogenetic placement software to say, "alright, I know my hits are going into these clades that I'm

interested in, in my phylogenetic tree". Or they're going into the outgroup and we can throw those away. Or they're going into something closely related that we're not interested in and we can throw them away.

What else? I mean, we do a lot of statistical analysis and we have before used machine learning. So anything in some of these different R packages where it's not the more advanced like convolutional neural networks or anything like that, but, you know, like boosted regression models, random forests, machine learning algorithms that are, yeah, like I said, they're kind of supplanting your standard regression analysis. So, you can get predictions but doing it in a way that doesn't have the typical assumptions that sort of parametric statistics requires, but also parametric, you know, standard parametric statistics. So, yeah, a fairly large variety of those types of software.

Interviewer: And you said you install some of them on the server?

Participant: Yeah.

Interviewer: When you work, for example, with an R package or a model, do you work on your local computer or do you work or where do you store the data and where do you work with it?

Participant: It depends. I mean, because it's mixed, really. If it's the large data, that will be on the server because you can't be going back and forth and downloading, uploading and all that stuff. So, for really heavy types of analyses that take time and you need to have a lot of parallel processes going on, we're always going to be on the grid for that. Or if it's going to be some sort of, maybe the dataset's not huge, but the computation in and of itself might take a long time. I mean, phylogenies are notorious for that if you're using a more advanced phylogenetic software. But then you could also be doing machine learning stuff where you because typically with machine learning, you have to tune the algorithm. And that means you have to put in ranges of different tuning parameters, and that might take a long time to process through to go through like a huge, you know, whatever n-dimensional matrix of different parameters to try to get the optimal performance that you're looking for based off of whatever validation method you're using.

Interviewer: When you collaborate around the data, do you use, what kind of platform do you use to share data or to collaborate on analyses? Do you work on the server or do you send? How do you?

Participant: Yeah, we work on the server. We're on the server. Basically, anyone who's working heavily in a project that's bioinformatics within the department. The way it works is that you give the system administrator a project name and then they make an account with that project name and you specify the members of that project so that they're the ones that have access to the data and can read, write in the different folders. Yeah. I would say that's the main way in which we would collaborate. If it's someone outside of the department, which isn't too, too common. I mean, with these projects, we really haven't worked(--) The collaborators, like I said, in that case, they tend to give more scientific input, theoretical input. And then we might share like a more finalized analysis that is not a very large file, for example.

Interviewer: Okay.

Participant: Graphs, you know, tables, things like that.

Interviewer: And how do you share that? Do you send it on email or?

Participant: Yeah, it'll usually be either part of the manuscript or it might be, [hesitates] but that's usually at the end. I mean, because we do put stuff up in like Zenodo or Dryad, but that's usually, like, right when the manuscript is about done.

Interviewer: Okay.

Participant: I guess we could, wait wait wait, we could also say Teams, we've used Teams. We'll have a folder for Teams, in Teams we'll have a folder for a particular research project.

Interviewer: Yeah [notetaker] you should also write down Zenodo and (.)

Notetaker: Yeah.

Interviewer: If I can read you, if I can, this is so small.

Participant: Yeah.

Notetaker: Not yet.

Interviewer: No yeah, thank you.

Notetaker: Zenodo then,

Participant: Write faster! [laughs]

Notetaker: Zenodo then comes in probably more at the end when you, kind of as a reporting step right?

Participant: Yeah, I was just going to suggest that you probably put that in you should put that in in report.

Interviewer: Okay or making, okay this is also part of making data available so

Participant: Yeah (…) and I can say that when it comes to availability, we've, we had have people cite these papers or they've actually gone into the repository and gotten kind of the curated information that we use to make our report and then they utilized that in their study and cited it.

Interviewer: Oh, that's interesting.

Participant: Yeah.

Interviewer: How did you notice that, did you, because you read their paper or?

Participant: Because they cited it. [laughs]

Interviewer: Yeah, okay. [chuckles]

Participant: And we could see the data that they used. [chuckles]

Interviewer: Oh, okay.

Participant: So that was nice actually.

Interviewer: Did you say Dryad also or?

Participant: Yeah, we've, we put stuff up in Dryad as well.

Interviewer: Okay. Just to go back a little bit to the analysis phase, did you encounter any problems or difficulties in, during analysis or?

Participant: That feels like such an open-ended question. [laughs]

Interviewer: [laughs] It's very open.

Participant: Yeah, yeah, well I mean I guess, I mean are you looking for a specific problem, or?

Interviewer: No.

Participant: Just the usual things.

Interviewer: In general, if there are difficulties regarding like the technical environment?

Participant: Yeah, because when you go to use analysis software that's being put out there as you know, public use, open-source research it's rarely ever going to be perfectly well done. It's always done in a hurry to try to get the publication out so there's always going to be stuff that's broken from time to time, or when they made the software they used specific dependencies, libraries and then those get updated then the software breaks and then you got to fix it. Again, this is one of those that just goes along with the territory, so it's always going to be that you know stuff breaks or isn't working quite well. Then if you can't fix it yourself the only option is to email the authors of the algorithms and of the software and say "look what's going on here" or you go on to whatever forums that you have. Because usually for more widely used software there are forums that you can go on, Google forums for example. I'll give you an example there's a popular sequence searching and clustering algorithm their program called VSEARCH, [notetaker] you probably remember that one.

Notetaker: Yep.

Participant: Yep, I was having a hell of a time with I think it was, it was one step that it does for chimera checking a specific type of analysis and it turns out there was a bug, in the software, because it was just spitting out kind of the wrong information or just

crashing and segmentation faulting and I just had to email the guy and ask “hey what’s going on” and then they fixed it and put out a new version within a day or two. And that’s a good experience because they were very quick and they’re very determined to make sure that the software works. Sometimes you might not hear from people.

Interviewer: Do you also write your own software or scripts?

Participant: I don’t, I mean… [hesitates] Scripts yes, but not what I would call software. It’s more you’re taking different software and sort of gluing it together with a little bit of editing within the script, editing meaning that you might want to be adjusting aspects of the sequence files or changing formats or making sure that you know feeding things through like a pipeline kind of structure. I dabbled in it and then I decided that I wouldn’t want to do that because then you have to be online checking who is using your software and fixing it constantly. I don’t do that and really no one else in our group at least does that but there is a [person] in the department that(--) written own software and it’s up and available public use.

Notetaker: But then you use Python or what kind of languages are you using in order to you know just fixing and combining stuff and like that?

Participant: It’s a combination of Python and basically Bash scripting. I have also been using more recently Nextflow, which is really nice. I recently constructed a pipeline that we use in the grid so that when we want to do our sequence clustering, you can choose from different options.

Interviewer: So, have you also made available your code or scripts when you’re publishing papers?

Participant: Yeah. So that’ll be up in Zenodo quite often.

Interviewer: Do you use GitHub?

Participant: We just recently got into that actually. We’ll probably be using that more we haven’t before. Although a paper that I was collaborating with outside with a group in

[country], yeah, that was going all that was going up on a that code was going up into to a GitHub that they arranged.

Interviewer: Have you encountered any difficulties or issues with making the data available for publishing data related articles?

Participant: I really can't think of any difficulties (…) Yeah, I don't recall anything that was particularly irritating besides just the usual, alright, NCBI once again changes upload instructions on how you put your sequences into the repository [said ironically]. Let's take some time and figure that out, which, you know, I have so much time to do [said ironically]. But, yeah, that kind of stuff happens all the time as well.

Notetaker: But when–

Participant: I wouldn't say it's a–

Notetaker: Sorry, just go ahead.

Participant: No, I was gonna say I wouldn't call it a problem, it's just annoying.

Interviewer: You can definitely mention things that are annoying as well.

Participant: Yeah, because it just takes more time to be like "oh, God okay let's just figure this out, do it", and then it gets done.

Notetaker: I was just wondering, with regard to this specific project where you actually downloaded the data. So, you weren't actually, you weren't needed to upload any data, but because you didn't produce any other data. It was more, then, citing the actual, the reused data, right?

Participant: Yeah. Yeah (.) Especially because you'll have, you'll have a lot of data that… Of course, there's a paper that's gone along with it and that goes naturally into the citation list. Any data in which there is no publication associated with, I mean, usually those are the ones that you got to be really careful with, and if we did keep that. I mean the accession number is there and that's the extent of the citation that's required. If there's no paper there's no paper but if people have the accession number and then there's some, like let's say it's like the most rudimentary metadata, okay, it came from

this lake or from that part of the ocean then we might use that to be like alright, we can classify it safely as this type of sample or this type of biome and then we'll use it. But for more detailed metadata there's 99 time, 99 percent of the time there's a publication associated with it and they get cited.

Interviewer: Have you had any costs related to making the data available or article publishing?

Participant: I don't think so. That I can't really recall I don't because there is Zenodo and Dryad are free and then of course NCBI if we're generating data that's free [hesitates]. No, I don't recall any costs.

Notetaker: When you, sorry(--) When you think(--) I mean in some way costs are involved when you(--) Because when you upload the sequence data and you(--) with that you upload a certain amount of metadata regarding the sequence data but then like when you do more environmental studies then you, at some point you kind of upload or you include that type of data in the in the actual publication. So, that there and obviously there's a cost involved there.

Participant: Yeah, yeah. I was just thinking that a little bit because… Yeah, if you're, if it has to be open source and you got to pay a little extra for the publication to do that, yeah, that can be a cost. But I wasn't thinking of that specifically with the data itself. It's more of the publication cost, I guess. I don't know if that qualifies as reporting. Because, yeah, that's always going to be a little bit more on the tricky side when you're required to do open source or journal fees, that kind of thing. That still comes kind of at the end.

Interviewer: So, in what ways do you report this, did you report this project? It's in articles, and are you also, do you report back to the funder somehow?

Participant: Yeah, I mean there's(--) Any of the funding agencies, they want like a yearly report, so you just simply describe what it is. And for example, for [funder] at the very end of the project you have to give a scientific report not just what you spent, what did you produce. Oftentimes those kinds of details will be given in whatever the agency is requiring.

Interviewer: Do they require anything regarding data how you manage the data or the data produced?

Participant: The agencies no. [university] does, I know that, because I because when I had my [funder] project granted a year and a half ago, I had to make up a data management plan, right. So that was required.

Interviewer: Okay, so that was in the planning phase or what did you do that later?

Participant: I guess you could say it was in the planning phase because a lot of the planning was in the grant and then when it was granted I was informed that "hey you gotta, you gotta make a data management plan" it's like yeah, okay. And the thing is with the management plan it(--) I basically described how everything is being backed up in the department because, like I said, with data that we're generating we have routines in the department on where raw data goes, the way the backup system works, again project folders. Because our IT guy kind of wants to, because we're legally required to make sure that we have everything backed up properly. The IT guy was adamant that we have to have a plan for every project to make sure that we meet the legal qualifications.

Interviewer: So, did you write that plan yourself? Did you ask(--) Did you use any tools for creating that plan or did you use any support function?

Participant: There was a template. Yeah, there was a I think there's, as I recall, there was a template provided like a questionnaire or a series of fields to fill in and I utilized that that was provided by [university] (.) if I remember correctly.

Interviewer: We haven't talked anything about the Preservation, how do you think about the long-term preservation of the data that you've generated in this project or used in this project?

Participant: I mean, again, this comes down to making sure that we're going(--) communicating properly with the IT support in the department. That the work that's been done and the data that we're using is backed up and once the project is completed that we archive it properly. And by archiving that means essentially making compressed

folders to save as much space as we can and then doing, you know, checksums on the compression, on the folders to make sure that there's no data corruption. If we transfer from one place to another, if they have to upgrade the server and we got to transfer from an old server to a new server that we're doing it in a checksum way. But that's kind of routine, I would say, in the department.

Interviewer: And those folders, they are stored on the department servers?

Participant: Yeah. There's actually a large, like within the [unit at the university] itself, there's a large data storage, and it's shared by a couple different departments. I think this is where the difficulty comes in between how things are done in this building versus kind of the vision for what [university] sees for data management in some way, because we're working with very large datasets. And I guess the example I'm thinking of is you have other departments, for example, they want to be completely Wi-Fi. Like, everything needs to be on Wi-Fi because of whatever reason and it's just that would not work here because of the size of the datasets.

Interviewer: That was actually all the questions that I had or do you have(--) Wait, I remember one more question. I was just going to ask you if you had any costs or issues problems regarding the preservation, I forgot to ask about that.

Participant: No. I see I mean the only cost, like, I said that because we have a department resource as I described the costs come from the overheads because you know as you're feeding into the department this is what this is part of the environment that we're able to work in. It'd be just like, okay, you're going into the lab and then you have all your consumables but then for the data you got to make sure that you have everything backed up properly, meeting the legal requirements for having things backed up and then also for the computational needs.

Interviewer: If you look at all the things that we have written, I don't know if you had time to read. Does it look correct? Or is there something you would like to change or add?

Participant: No. That I can really see. No, I think that, think it's okay.

Interviewer: Okay. I would like to ask you about these figures up in the corner. If you were to say how you feel about the planning phase, if it worked well for you or neutral or if you feel dissatisfied.

Participant: Yeah, I mean, I've been pretty happy with, again, to limit it to my experience working in this department, I'm pretty happy. Because they, I mean, you know, everyone's busy, and it takes a bit of time, it could be a bit slow, but I'm pretty happy. I would say that's alright.

Interviewer: And data collection. How do you feel about this?

Participant: The internet speed could be higher. So, let's put a middle one there.

Interviewer: Yeah, okay. The analysis phase, how do you feel about it?

Participant: I've been happy with that because we have the resources we do.

Interviewer: And making the data available?

Participant: Yeah, I haven't. I would say that's fine.

Interviewer: The preservation, are you satisfied with that?

Participant: Yeah, that should be okay.

Interviewer: And reporting, writing your articles and reporting to the funder?

Participant: I don't know. I guess that's all. I haven't really had problems with that, except for the occasional reviewer. [laughs]

Notetaker: Not in the e-infrastructure. [laughs]

Participant: Yeah, not in the e-infrastructure.

Interviewer: Okay. Finally, I would like you to describe your dream scenario. It can be big or small things. If things worked perfectly, what would that look like? Are there things you wish for?

Participant: If things work perfectly, [university] would have like a large computing grid, which I really think they should. This has been something that's kind of caught in my craw because we basically have, our department has a really good grid. And of course, I think it's kind of rare that departments have that kind of resource, at least as far as I know at [university]. UPPMAX was working great but then as soon as NAISS came along and then we have to start sticking to Dardel, that's been a bit boring because it's a real pain in the ass to use. So, you talk about a dream, I really think if [university] had, like, a nice consolidated one one-stop shop for storage and grid computing scientifically, I think it would actually be it's something I think the university really should do, and I'm not alone in that.

Interviewer: Okay. Is there anything else you would like to add?

Participant: I don't think so.

Interviewer: [Notetaker], do you have anything, any additional questions?

Notetaker: Nope.

Interviewer: No? [chuckles]

Participant: I hope my answers made sense. It's Friday and you know, the brain stopped working about a half hour ago but...

Interviewer: No, it's very useful information, very valuable. So, thank you and you're also very welcome to contact us after if you have anything you would like to add or also if you have any questions regarding data management and some of the things we discussed today you can also contact [university data support function].

Participant: I'll tell you what, if you guys can just keep pushing to have a centralized grid computing facility and storage facility that'll be, that'll make my day. [laughs]

Interviewer: Okay, yeah, well [chuckles]

Participant: Okay.

Interviewer: Thank you so much.

Notetaker: Thanks a million.

Participant: Yeah, I hope it was helpful and again nice to see you again. [Notetaker] it's been a while.

Notetaker: Yeah, let's have some lunch.

Participant: Yeah, I could do that.

Notetaker: Thanks a million.

Participant: No problem.

Notetaker: Have a nice weekend.

Participant: You too.

Interviewer: Thank you.

Participant: You too, enjoy the weather while it lasts, supposed to be crappy this weekend.

Interviewer: Looks nice now.

Participant: Yeah, it is okay.

Notetaker: Bye bye.

Interviewer: Bye bye.

Participant: Bye bye.