

1 **Interviewer:** Let's talk about this picture that you see now and about the individual  
2 phases of the data life cycle. Starting with the first phase, the planning phase, how does  
3 that work for you and what do you do with data during the planning phase?

4 **Participant:** When I use data, when I need data to make a scientific point, this is often  
5 the goal, or, in a few cases, we could also have some strategic goals of realizing a way  
6 to collaborate by having a place to dynamically upload, or as the data comes, we will  
7 upload it and our collaborators, specific collaborators, can then follow what we're doing,  
8 how far we're progressing. But generally speaking, the planning stage would be  
9 something that we either set up in connection with an application, or it could be that we  
10 realize that we need a sufficiently large set of data, which are in my case calculation or  
11 results, to make a certain point and admittedly we often find that "oops there was  
12 something here we hadn't thought about" and then we have to add that. But generally  
13 speaking, we can sometimes succeed at making a plan for proposal, or plan for a paper,  
14 where we more or less delineate what is needed to succeed. And I think that is a stage  
15 where I don't think I need so much help, because it's quite individual for the kind of  
16 papers that we are aiming for. So I don't think I have too much to say here, actually. You  
17 know, it (--) that's it. Okay so this is the kind of papers I write. You could say that if we  
18 were to have a collaborative effort, for example at [university], facilitated by data that is  
19 collecting what we, you know the thing we pull together, then we would probably need  
20 some help of getting it in there efficiently and getting it summarized efficiently. And also  
21 training of our collaborators to quickly see that (--) what the metadata is. And I know that  
22 there are some systems like that, but this is more on the later stage when we actually  
23 create and collect the data, I would say, than here in the plan. But when you plan, it's  
24 also good to know, to make an agreement with collaborators, "this is how we would want  
25 to do this". So there's somewhere between the planning and creational, the collection  
26 state, I would say. For these tools can come in. But the individual planning, does work.  
27 You know it's on the researcher's initiative, I feel, I could learn more but I think that that's  
28 what we (--) that I think is not the problem at least for us

29 **Interviewer:** I understand. Then we can directly continue with the next phase, the create  
30 and collect phase. How do you create and collect your data?

**Participant:** Well, it's contingent on how much data storage I have available and how much computing resources I have available. But one can say that, you know, in the worst case of scenario, I can only keep like the input file for calculations and the ASCII output file. And then, basically letting that be a tripwire for when people want detailed analysis of a particular sample of the result. You know, for a particular figure, we might then recreate the data, the full set of data that the calculations actually give, and then provide the analysis for that. And that is something, that's the minimal thing we have to store. Right now, it is basically ASCII in and ASCII out, basically. And I personally use either a bash script, sometimes some Python to steer it with. Some people like to use a graphical user interface for controlling that data. I think this is going to be (.) To make a recommendation at that stage, I don't think it's part of, basically the data. But we have some ways of doing this.

I also know that there are tools so that multiple people can put it in, and as long as people have agreed on what metadata and what kind of standard there is for these calculational data, there's a database tool that can sort of collect it and make that available for each research and so on. There, however, I don't think it's just the tools. I think perhaps some educations of [university] researchers would be good because some of my friends who have been trying to use the computer data for machine learning, have had this nasty experience that unless everybody does exactly the same kind of quality when they create the data, machine learning tends to want to learn the differences between contributor A and contributor B, rather than actually extract some valuable information. So yes, we would like to have better tools! And that means that we all sit down and learn them and we are terribly bad at that as researchers! We don't, [laughs] we don't get, we don't create the time for this! And the other one is that we actually have some, perhaps training on using this together, so that one can sort of define standards and decide I'm sticking to those, so that our data are interchangeable. I hope this is helping? But this is a guideline that unfortunately we've learned it requires some self-discipline. It's not only the tool, it's also the discipline or self-discipline in sticking to that.

I think this is the overall thing. And then in practice, I use first principal density functional calculations, and I don't always know enough about how to extract the data, but that is,

61 in some sense not something I want to (--) I don't think you can help me with that,  
62 perhaps. I think it's actually my job to stay on top of what's possible. But when I want  
63 something, it would be great if you can help me get it out again of the calculations.

64 **Interviewer:** Okay. I want to ask about data management plans. Have you ever used  
65 one (--) do you usually in a project in the planning or create and collect phase?

66 **Participant:** It's a very good question. I did try actually once, because I was hoping this  
67 somehow would lead me to a way of storing. You know I have about 100 terabyte of  
68 data that (--) they're not all filled, but I need some swap space for actually doing  
69 calculations, because we temporarily create a lot of data per calculation, and we can't  
70 afford to store it all. So what I need on three different resources at [compute center] and  
71 then at the other computing center, we will soon only have one national center of course,  
72 is a lot of extra space of either 20 or 40 terabytes of free data so that when you get an  
73 idea, you can sort of follow that through and perhaps harvest everything. What I would  
74 love there to be, is something that just summarizes that main result and keep track of all  
75 the metadata, so I don't have to do it. I know this exists, but I haven't learned it yet.

76 I filled out one of these data management plans and I felt there was something (--) a lot  
77 about how do I keep it secret, how do I do this and that whatever. And I actually think  
78 that if I go to (--) I believe that [institution at university] has a database for doing this and  
79 I'm sure that there's others. I work with Quantum Espresso which is a [inaudible] suite  
80 and they also have a separate model (--) have a tool for workflows, that would also be  
81 nice to store. This is something that I'm terribly bad at, I would like to be helped with this!  
82 [laughs] but I don't actually get there! So I don't actually use the data management plan  
83 and it's all up to me to remember where I placed it. So I basically work with the libraries,  
84 that reminds me what was the input file and (.) that's not good! Because, let's say I got ill  
85 somewhere, so nobody could use this data, they wouldn't know precisely what the logic  
86 of this subdirectory was. So that's something I could really learn!

87 **Interviewer:** Yeah, so if I understand you correctly, you tried using a data management  
88 plan and put some value in it, but in practice, it didn't turn out to be useful for your  
89 specific case, right?

**Participant:** Yeah, and probably because, exactly (--) So the thing that was, it was a questionnaire, and I never got anywhere near, basically, a database management tool, nor did I get anywhere near (--)(.) A lot of the work has to be done also by the researchers in figuring out how you structure it. Do you say that soda potentials, in my case, soda potentials, is that more important than type of functionals, type of problems that you're dealing with? So yes, and it's hard. And then, complicating with that, and none of the questionnaires got me anywhere close to picking a tool for doing this, so there is something I can get help with. But complicating with this is also that (--) although it sounds crazy, we really really would like to have the same cutoff everywhere so that we can eventually machine learn them. And I also really really would like to have some kind of process, some kind of idea of how much space I have, so I could extract as much of the relevant information while I have all the data, and then compress it down to, things that are just being preserved. I can always regenerate it, but it would be nice to have extracted the common questions about system X, Y, and Z in various functionals in various methods, in various things. And always having the same quality of the calculation, at least have a metadata that sorts them into is this, type a, type b and c. Because otherwise I don't want to mix it in and actually try to do any machine learning on top. This is things I really really need to help with, it's personally all being down in my head right now.

**Interviewer:** Yes I understand. Okay let's continue with the analysis. So you've created your data, how do you work with the data analysis then?

**Participant:** I'm probably old-fashioned but one of the things I feel that the first thing I need to do is to plot, whatever I can do because that's part of the story. So unless I have it plotted in some way that shows some systematic variation or dependency or correlation, I'm not even grasping the actual research story yet. That too would be different if you were doing (.) a social science database, of course, and you don't want to have pre-constructed notions about what the story is you want to tell. But I do actually, because I'm not necessarily interested in, let's say, just hunt for new and better materials. I'm would be interested in knowing why is this happening and if I'm not actually succeeding in capturing the physics that I want to have, you know I test whether

120 I capture the sufficient physics, well then I know why is it failing me, and I can go back  
121 and try to do better on the way I define the calculations, the quality if you want of the  
122 calculation, so in my case develop the theory so that it can actually capture more  
123 physics.

124 So I ask questions interactively with this data. So there it's helpful to have the metadata  
125 telling me, it's very useful to have all data that I have created, or my friends have created,  
126 at the exact same quality so I don't make false conclusions. I can actually compare it.  
127 And that is one of the reasons why I haven't actually chosen a structure that can  
128 summarize this for me. Some of my former graduate students has been helping me, so  
129 for particular problems I've actually managed to do this. But generally speaking I haven't  
130 found one solution all that fits all. But once I have that I can start asking the scientific  
131 questions or if I can just pick. I know how I organized even this sort of file systems, I  
132 know that in that file system there is potential answers to this effect and I can go test it.  
133 And if I'm missing data then i know that to create more data. But the analysis is basically  
134 to me part of the story. It's the story that I want to tell like I slice up, let's basically the  
135 data, or I slice up the question, so that I can extract some conclusions don't know if  
136 that's easy to automate not sure about that, because you've got to get a good idea and  
137 you've got to get an idea nobody else has thought of, right, and you want to publish it so  
138 I'm not sure there's a prescription there [laughs]

139 **Interviewer:** The tools that you're using there's Python or something similar?

140 **Participant:** Yes it's often Python or it can just be shell script or (.) as (--) So, you know  
141 you get graduate students and they're younger than you and so they complain that I  
142 sometimes use Python which really just calls up bash scripts that I've already made. So  
143 it's a mismatch [laughs]. I was actually just contacting the [name of research group] and  
144 they actually have realized this problem a **long** time ago. This is Python but they keep  
145 track of workflows because some of these workflows are indeed recyclable. I'm having a  
146 hard time moving off bash because I know I had a solution there. It's a little bit like if you  
147 have a problem you google for a good code suggestion. And then you see you're  
148 adapted to your particular specific needs for data extraction because you have this idea  
149 that you want to, I say, see if it's true. And it takes too long effort to actually make me, to

150 make a common platform, I don't have the staff to do this. And sister group, the  
151 [research group] and, of course by obviously the [research group], they actually have EU  
152 funds to set up workflows and so on. But as they also acknowledge, that takes you a  
153 little bit further away from the "oh I get a good idea I want to know this that or whatever".  
154 So in any case I would like to have both those automatic tools and maintain a good  
155 knowledge of how I can always hack something in Python or bash. I'm a little bit of (.) a  
156 you know retro researcher like that but it's also, it's a time issue.

157 **Interviewer:** Yes. Anything more you want to mention maybe about problems that you  
158 face in this phase?

159 **Participant:** Well most of the time when I have problems it's because I forgot to actually  
160 ask part of the key question right [laughs] I realized there's something and so it feeds  
161 back on the data creation part. I think some kind of tool that's that that will say you know.  
162 Let's say that I only have 50 of those kinds of calculations. Right now, it goes by memory  
163 If I have some kind of database management systems, I could probably find, "oh, okay, I  
164 only have 50. That's not nearly enough to draw any sort of systematic conclusions". So,  
165 again, here, I could learn to be better even at this stage. And (.) I would still have the  
166 other tools for this, you know, fast thinking, but this way would be faster to figure out  
167 whether I need to work harder.

168 **Interviewer:** Yes. How about the next phase, making data available? That means  
169 sharing internally while working on the project but possibly also sharing the data in  
170 general.

171 **Participant:** Okay I receive an F on that [laughs ironically]. And it's even to the point  
172 where you know (..) well I mean I'm asked because some of this data that (--) Some of  
173 the conclusions that I've drawn, for example in a recent paper, actually was to show not  
174 only could we do something that we didn't first think we could and I don't think  
175 everybody thought you could do from regular DFT calculations, you could sort of all of a  
176 sudden convert them into something much better. Or (.) you extract more information  
177 than you thought you could before, in this case about Cauchy particles. And then to  
178 make (--) I wanted to make the point then (.) that you are doing this without changing the  
179 underlying description of the structure, so that the conclusions that you're drawing on

180 these light-matter interaction processes actually is pertaining to the good description that  
181 you already had of the structure.

182 That meant proving this across very many different assumptions. And the easiest way  
183 was (--) to do all the calculation first and then going and then taking out the slices. And  
184 then after that people said "oh you should publish this as a separate part" and I don't  
185 know how to do this [laughs]. I know that there are real reasons to do it. And so right  
186 now I have (--) this this is basically 40 terabytes or something like this of data that's just  
187 lying around. And I try to figure out how to do this but I don't know who has to pay for  
188 this. I believe I asked actually [e-infrastructure] what this would cost, just to have it  
189 stored because of this. But maybe there are other solutions, we are working on that. But  
190 you can see that (--) this is one of the things that has been lost in the in the flow of many  
191 things that we also need to do, right? We have postdocs and graduate students and  
192 master students that need to be finished right? So here would be something where (--) it  
193 may just be as simple as knowledge and we can sort of lift it off the servers at the [e-  
194 infrastructure] center and onto some national facility or something like this. But I don't  
195 know what or even what it's going to cost. So, you know, I don't, I'm not sure that, I don't  
196 know, it's not obvious that (--) depending on what it costs, that I could actually do this.

197 You may laugh at this, I'm not sure this should go into the transcription, but I've actually  
198 bought big terabyte data here in my office, so I have a plan B, until I figure this out. It  
199 was cheaper to buy a 20 terabyte hard disk than to sit down and figure this out, right? F  
200 is the grade that I get, right? [smiles]

201 **Interviewer:** But you do have projects with collaborators where I assume you do some  
202 sort of data sharing?

203 **Participant:** Yes I do! And there too is probably an E because I have put this up on the  
204 [university] git so that my collaborators can see how far I've progressed and you know  
205 how good are their collaborations. And also, the thinking is that I ask my collaborators to  
206 sort of agree that my set of collaborators (--) I don't want to have to say that "you can  
207 only see this or you can only said that", you know. I tell them just to be respectful of the  
208 other collaborators I have. And that tends to work. That is one data management plan I  
209 guess. You know, I just asked for respect among students and among my

210 experimentalist friends. But generally speaking, then they can see, sort of some of the  
211 other things that we can also do now that we're at it.

212 **Interviewer:** The main issue with git that I understood was the (--) that you essentially,  
213 allowing (--) giving everyone the same rights?

214 **Participant:** I have split some of it up in soft directories because there is one thing that  
215 involves, some software involved, and they have a contract. But generally speaking (--) I  
216 actually think it's good that they see each other [laughs] right? I really think that [e-  
217 infrastrucure] or me for that matter, Git could actually help by allowing them to speak,  
218 because they can't do much if they just see the numbers right? I hope that you will, you  
219 know, help us collaborate better, it could be useful [smiles].

220 **Interviewer:** We'll do our best [smiles] (.) Okay there was one point in the make  
221 available that is not in the notes and that was that, for sharing data, you usually use git?

222 **Participant:** Yes and I would like something better [laughs] Because git (--) the git wiki  
223 is not safe right? There's no (.) Well someone told me that there is a way that, I can go  
224 back and see, basically like openly if I can see what has been changed and so on. But  
225 something better than Git for making sort of an overview of data and getting inspired to  
226 plot something exciting would be good. So I'll do the Git also so everybody can follow.  
227 But (.) maybe it could be automated, right? If I tell you how to (--) Basically I use grep or  
228 bash scripts and then grep for whatever property I want to get out because then I have a  
229 record of what it was I extracted whether it was with or without electrostatic decoupling,  
230 for example, it's a very technical thing. But if I don't do it systematically and if I don't  
231 keep a record of it ,it's lost right? So something that would convert something into even  
232 just the basic tables or something like this. And I suspect that these tools exist. I just  
233 haven't learned them. I'm giving myself an E on that part.

234 **Interviewer:** Good. Shall we continue to the data preservation and archiving part?

235 **Participant:** Right. Yes. I'm also here a loss. I think I told you, whether it was off record  
236 or on record, but it's fine if it's goes on the record. I, you know (.) there is this issue  
237 about (--) but there's actually two issues. In principle I could always recreate everything  
238 as long as I have the input files. And they don't take much, I mean, I can have that on

my personal computer I actually probably have it somewhere. Because that's the whole point about these first principle method is that you just you know you have your initial guess and if you're really luxurious you have the final guesses, or maybe you started a little bit different but you could in principle contain everything in a very small memory package. But that's not good! because that also means that you're sort of completely self-observed about, you know "I just wanted my own points" and the point about data for collaborations would be that other people want to understand why it was that this system is bound more strongly than this other system. And for that you need to go back and do more analysis so that means that I can either rely on a tripwire sort of systems, where I train them to ask me, you know a month in advance so I can recreate the data by the time they need it, or more likely that I learn which is what I do now, from experience that they are likely going to ask me about this question. Or this is a system that I know somebody locally is working on and therefore I do a little bit more analysis before I throw away the core XML file that contains all the all, the real hard numbers right.

So this is what I do now, so this is the trade-off I have that I pit myself somewhere in the middle. For the most likely system, I already have some analysis ready for them. But the system is not perfect. I would love to, you know, petabyte is the thing that would be great. Then I would have everything within fifteen, within very short time. Post-processing to get this kind of information could be extremely fast. I don't know how to solve this! Because I don't have the money and I don't think any place will give me petabyte of data, from the outside, saying that somebody at [university] or [country] want to ask this particular calculation that I did like in 2002 for detailed analysis right? There's a cut, there's a balance. And don't know how to do this. Preservation to me is like I just try to learn what people can ask me for, and I make mistakes, I save too many of some things and not enough of other things. But it's not clear what I should save, to facilitate fast interactions.

As long as the papers, these specific papers are ongoing, it's not a problem. I just keep everything, right? But then when the paper is out, you know, how many more years can

268 you hope for somebody to come back to you and ask for a detailed analysis of this, that  
269 or whatever?

270 **Interviewer:** And do you do preservation on your personal hard drive or?

271 **Participant:** Let me put it this way, as every year I create more data, so the things that I  
272 think that somebody (--) I hope that somebody will be asking me about, I actually keep  
273 there, sort of like half of the stories I have so that they are like ready, on-demand-  
274 information if you want. And I have the other part, which is, so that when it's ongoing I  
275 definitely need to be like, I need to store everything right. And then I have my local hard  
276 drive I dump everything that's like input files.

277 And I normally do this once a year, which is, of course, way too little [laughs] in terms of  
278 safety. That would be good to automate that also. We could probably find the list of files  
279 that these are the files that would allow me to recreate everything. And I could probably  
280 do that, actually. That would be part of the preservation, that would be a better way of  
281 preserving it so that, you know, it wouldn't take more than a terabyte to save everything  
282 I've ever done and all the code work I've ever done, basically. That I could always afford  
283 of course! Both are copied locally and I'll pay for it over at [e-infrastructure] right? I mean  
284 terabytes now (--) I think even I think even that [university] we have a terabyte  
285 somewhere? So that minimal storage, that's of like this is my life kind of data set, that's  
286 just a question of me actually putting down and doing it. But for those on demand things,  
287 you know I perhaps need some advice but it's also (.) yes it's not easy.

288 **Interviewer:** Yes, I understand okay let's talk about a reporting phase, how does that go  
289 for you and how do you report?

290 **Participant:** Well I have troubles with referees occasionally but I think we all do, if that's  
291 what you mean. The reporting is of course also to funding agencies. But my data is  
292 something that I actually think transcends, I know it's a weird word, but the individual  
293 grants. It's actually data for a program rather than data for a project, like a funded grant  
294 or even a subproject like a paper. That's why I really think should we should have some  
295 automated systems that or everybody at least that does a certain class of studies or  
296 something like that should identify what is the minimal thing that we need to recreate it.

297 Admittedly and a lot of cpu hours, but at least recreate what we already have and save  
298 that. That should actually be something we should be doing because it's unsafe and try  
299 on horse will kill us right. That I think is actually, you know, it should be, you should just  
300 keep it basically. And with some help on making the structure readable even after 10  
301 years or 20 years. I have projects that I haven't looked at for 20 years, but they are still  
302 data.

303 I don't really know what you mean by reporting. The reporting on a paper is often not a  
304 problem because that's when you've been active in actually analyzing them, making  
305 these sort of, "Oh, I can plot this in a new way that actually reveals something that I  
306 hadn't thought about before" and hopefully this is good for everybody that you know I  
307 add my perspective of how this should be understood in a paper or for that matter  
308 reporting to SSF or whatever. That's also, you know that's part of the recording. But the  
309 data that we create and, how do I say this, particularly with first principles calculations  
310 that some of us here at [university], that we actually have quite a lot of people at  
311 [university] doing this. Then that data is, I mean, it's not going to stop being true 50  
312 years later [laughs] right? It is, it's the core of insight. It needs to be in our head so we  
313 can sort of synthesize it, we put it in textbooks, but it also needs to be somewhere so  
314 that you can, you know, open that box again and expand it.

315 Sometimes, yes, it turns out that you can do it much faster today. That's wonderful news!  
316 You know, it's getting technical, but for example, if one of the things we can do is to do  
317 structural relaxations when we first started, we couldn't even do structural relaxation, so  
318 we had to save loads and loads of different configurations. This data is obviously  
319 obsolete. But if you just keep the input data, it's, you know I don't even see it on my hard  
320 drive anymore because 20 years ago it wasn't very much data. So just if we just cut it  
321 down, we could keep the smallest part and we should actually be doing this. And I don't  
322 even think that this is a problem of preserving or reporting to, you know the next  
323 generation for that matter. I mean there's a (--) what I'm trying to say is reporting is not  
324 just to some extent external people but it's also to those who want to see this later. Our  
325 students, graduate students, can learn from each other that way. If you talk about  
326 reporting in terms of, which is something we underdo in my group and I know other

people at [university] are better at, would be that we kind of do papers that are like method papers and then we do papers that are result papers. And the method papers is often tied to a code and then we release it often in open source so that other people, there's a great synergy in that. Sometimes we want to do it in a closed code, just because, well this is a difference between chemists and material researchers historically. you know. But this kind of reporting is also it's also worthwhile to invest in because it generates (.) like say interest long-term, and there we really would like to learn how to put this up on some national service so other people can make use of it, and I think it's a question of sitting down and learning this.

**Interviewer:** Are you talking now about the software?

**Participant:** Well the software for sure that just goes into those (--) we have these open source software for sure. So I just report it. There's some git system, I've forgotten how you do it by the time I get to do it again. But the other people help me. I upload my functional so that they're available in other people's calculations. That's taken care of automatically and you (--) Great you get a lot of testing that way because loads of other people you ever thought would use it are actually then using it all of a sudden! But then there's also (--) if I have this database that I created to show that "yes I could actually get not only structural data but at the same time causal particle information whatever that is – light matter interaction results - " then that would be good to place at least sort of the minimal information of that available somewhere that other people can look at. But with enough information that they can recreate it. And that is something that I am not doing enough of [smiles]. In fact that is, you know like lawyers which would talk about billable hours, but I think for researchers that's another way of actually creating let's say visible outcome of for example [university] activity. And so there I would also like some support or just somebody would give me time to learn how to do this [laughs]. Because you know that it's becoming almost as important as actually writing a result paper or writing a code paper. You know my group certainly does this way too little! So there I would love some help! I you know I try generally just to tell you where I'm bad because then hopefully you know where some people can be helped, yes the logic of what I'm saying right?

357 **Interviewer:** I would ask you now to distribute these three smiley faces that you see.  
358 One to each of the different faces and they should represent your general feeling how  
359 well this phase is going.

360 **Participant:** And not how much time (--) not how, you mean, okay so let me clarify this.  
361 There are some phases where I'm not very good at it but it's actually a lot of fun. But it's  
362 not it's not the fun factor you're after it's how I satisfied I am actually with the efficiency  
363 of my work that you're asking is that correct?

364 **Interviewer:** Yes rather that exactly.

365 **Participant:** Okay oh wow what a challenge! Okay so happy face let's see now could  
366 you move the window because I can't really (--) oh I should just get by the by the central  
367 circle here. (.) Bad, you know the reporting, you know thinking through the reporting and  
368 the preservation is probably a sad face because we're not there yet. And I think that we  
369 are not good at analysis but we make it up for by enthusiasm because you know  
370 researchers do succeed in creating papers anyway. So I think that in some sense, it's  
371 somewhere between a flat face and a happy face. And that's the analysis, right? I'm  
372 saying I could be a lot better, but we're obviously not bad as researchers. And I'm not  
373 talking, I mean, massive researchers in Sweden are not bad. We do create interesting  
374 papers.

375 The create and collect, I'm actually quite pleased with doing Bash and Python and so on.  
376 A little bit, you know, a little not perfectly smiling face would be on the fact that, you  
377 know, it's not always easy to have an overview. "Should I create more data?" And  
378 sometimes you end up having enormous amount of data on something where you  
379 already made the point as it turned out when you get to the analysis part. But generally  
380 speaking, the create and collect is working quite well. We'll put a happy face there, you  
381 know, just you know we you know one should not be I'm not saying we couldn't be better  
382 but it's actually something we manage to create a lot of data I think, that's for sure. And  
383 some results that can be published right. And the analysis is like you know where we  
384 could be better and actually helping ourselves in doing this creative stuff, maybe it  
385 wouldn't take so long. Don't know if it's just a technical, a technological limitation. It  
386 could also be that creativity takes you know we're not as creative as we wanted to be

387 [laughs], but at least having a tool that would sort of allow us to follow ideas more quickly  
388 could probably make us better. Make available preserve and report yes well if it's about  
389 actually storing the data in such a format that it's useful for ourselves and for others, and  
390 then it's probably the saddest of the faces.

391 **Interviewer:** Okay that was make available also sad face?

392 **Participant:** Yes well you see I can't really see what I said about over in the box what  
393 above the make available but. Yes I'm not good at making it available, that is a sad face.  
394 So you pretty much have the, you actually have the right alignment up there that (--) to  
395 the early stages I'm happy about, and then it gets to be the flat phase and then. So the  
396 order that you have the three happy phases up there on the slide is actually, pretty much  
397 like in this wheel where it goes basically.

398 **Interviewer:** Okay so the planning phase would you put a happy face there then?

399 **Participant:** Well I'd like to be more successful at getting grants [laughs] but I'm not sure  
400 that this is (.) I thought I only had one happy face and one flat face and one sad face but  
401 you're...

402 **Interviewer:** Sorry, no I would like one face (--) one of the smileys for each of them.

403 **Participant:** So we got create and collect was happy, was happy face as good as it  
404 could be. Analysis is probably a flat face I think it could be better, the sad face is for the  
405 rest of it. And then the planning (.) Well I'm always excited about applications where they  
406 go in right, so I think I would, with apologies, like to say that we are okay with that. But  
407 it's you know contingent on the observation that we rely on a lot about, you know,  
408 intuition and I'm not aware of some place where you could help us in any way. I don't  
409 know if I could help that.

410 **Interviewer:** Sounds good. My last question to you. You see there is an empty box in  
411 the top right corner which is a dream scenario. Please describe your dream scenario  
412 and that can be any big or small changes in the process of your research data  
413 management.

414 **Participant:** Oh, well, I need to think a little bit because I see this as the (..) You know,  
415 there's all these things about, it would be nice to find the time to automate some of these  
416 things. But I also see that since the last three of them were all bad, it's difficult to  
417 differentiate about, well, I feel that could be a lot better over the last three, say,  
418 components of the wheel.

419 **Interviewer:** And it can be several of those bullet points.

420 **Participant:** But what I would actually like to see is that we had some (..) Some of my  
421 collaborators actually have done it, that you work with this over many years. I've been  
422 working on functional designs, I've been trying to make DFT more accurate for many  
423 years. And the benefit of that is you don't make mistakes as often as you used to, you  
424 understand? And if you work with data over many years, I'm pretty sure also that you get  
425 better at the planning, you get better at the creativity analysis, you learn to slice it up in  
426 the beginning.

427 So I think a dream scenario, which is something we tried in a recent application, is to  
428 actually set up the entire chain, so that we have a database where you can go in and  
429 you could trust the data, because you have been checking it as it goes in. And you could  
430 trust the data to have a common quality so that you actually dare learn something from it.  
431 Now, it may just be that it's a targeted goal of optimizing material purposes for one  
432 particular function or one particular type of material. That may be a much more realistic  
433 thing, but having all the tools available at, you know, in working conditions so that you  
434 could actually build that and have it built up so that you could use it years over years  
435 over years. Done pretty much that for testing, for you have a setup for testing quality  
436 inscription of molecules, you know, molecular properties. I think that is actually very  
437 good. I would love to do it also for some of these more technological issues that pops up  
438 when you try to filter pollutants, if you try to make clean energy. There's lots of these  
439 problems that are related to these weak interactions that we focus on, but where I  
440 haven't had these long-term investments.

441 It's not enough to do 2,500 calculations; you need to have done millions of them,  
442 basically, to really have something that is of value to other people. And that takes a long  
443 time. And it takes also that you talk to people who know how to set up such a database

444 that is available to everybody. But if you do that you actually can really learn something  
445 from that. And so that is a dream scenario that you could somehow get help to (--) either  
446 you team up with somebody who has the let's say, the other part of happy faces. Or that  
447 you did you get somebody to help you to do not just one of the things but have the  
448 whole package so it really works and it's appreciated as the go-to spot for materials x y  
449 and z types for example, you want to challenge x y and z.

450 That is a dream scenario it requires a lot of thing it probably also requires a lot of (.) You  
451 know the people who have succeeded with this I know is at Berkeley or that I have some  
452 friends in Korea who succeeded with this, and it has been a deliberate strategy over  
453 very many years. You know it's not something you just build up. But we are a little bit too  
454 isolated to tackle all of those things, to be good at all of these things together, so some  
455 kind of platform is a wish list.

456 **Interviewer:** Okay that sounds like an interesting dream scenario.

457 **Participant:** It's a dream scenario I realize that. It costs money that's a lot of money and  
458 time.

459 **Interviewer:** Thanks a lot for your input do you have any more questions or things you  
460 want to mention before we close this?

461 **Participant:** I just want to say that I have (--) You know this is all for my major part of my  
462 research which is about trying to make, in my case, first principle simulations of material  
463 from, and for chemistry better. It probably varies, but interestingly enough I was recently  
464 asked to participate in work with industry who are also trying to, use their data, to  
465 optimize for example processes or manage risk and there's been all these ideas. And  
466 some of these things about actually controlling the quality of the data that goes in, is  
467 actually something that they are also struggling with. So I think that (--) I wish that we're  
468 not only collecting data, but we also have some kind of attention or, to have a common,  
469 agreed, quality limit on what this data should be so it can be compared. It's just a fear  
470 that I see that data is not good in itself. It needs to be (.) managed!

471 **Interviewer:** Absolutely, I understand.

472 **Participant:** It's just something I learned, basically. If enough people worry about it,  
473 maybe something will happen. That's why I'm adding it right?

474 **Interviewer:** Then I will stop the recording now.