

1 **Interviewer:** Okay, I will first talk a little bit about the project in itself and then I'll talk you
2 through the plan for this interview. So, the aim of this KAFFE project is to map how
3 researchers work with research data and what their needs are. The goal is to develop
4 the best possible support for you as a researcher, such that you can focus on your
5 research. And to structure the discussion, we use a simplified model, and you can see it
6 already in the screen sharing. It is based on the research data lifecycle, including
7 phases like planning, data collection, analysis, accessibility, sharing, preservation, and
8 reporting. In this interview, we are specifically interested in capabilities. And by this, I
9 mean the capacities of the system you work in. That's a very broad concept describing
10 processes or information resources that are available to you, also IT support, and so on.
11 So, in other words, all the support services that are needed to be able to manage
12 research data well through all these phases of the model. We also put special emphasis
13 on investigating e-infrastructures and what is needed to support you in the best possible
14 way. And we are aware that e-infrastructure is not something that you usually think
15 about in your work unless it is not there or it's not working.

16 So [notetaker] could you share slide one which gives an overview exactly over the
17 research structures? We're thinking in these four different levels here, first one is the IT
18 infrastructure, level two is the research infrastructure helping in data creation and
19 processing, level three is more like tools for working with the research data and level
20 four is actually people and information. So, it's, yeah (.) and please keep this image in
21 mind throughout the interview as it represents the environment that you relate to in your
22 daily work.

23 Today's interview goes like this: we start from a research activity that is relevant for you,
24 maybe something that you are currently working on or that is recently completed. And
25 we do this because it's usually easier to remember things if you're thinking about a
26 specific instance. However, if you have comments that are general, outside of this
27 project, feel absolutely free to just add them.

28 So, now we work with this, like I said, this simplified wheel containing the phases of the
29 data lifecycle. And now, [notetaker], could you show slide number two? Exactly. So this

30 contains a bit of these capabilities that I mentioned before, like, or different parts,
31 different things to think about in the research data lifecycle, which is questions like
32 defining the project's data, thinking through data management, creating a data
33 management plan, finding data, generating data, processing data, but also document
34 everything that you do with it, analyze the data, publishing, storing, archiving, and
35 reporting. So now when we will talk through your project, you don't need to really refer
36 to all of them, and just speak freely and we'll see how that matches the different phases.

37 Right. So, like I said, [notetaker] will take notes of what you're saying and while he's
38 writing you will see what is noted down and at the end, you'll be able to go through it
39 and sort of accept this version of the notes. Do you have any questions before we start?

40 **Participant:** No, I think the introduction was good. So, I tried to also orient myself
41 towards what you're interested in.

42 **Interviewer:** Great, thank you. So, let's start with a simple question without notetaking,
43 just to set the ground. Can you tell us very briefly about your research with a little bit of
44 focus on the data that you work with?

45 **Participant:** Yeah. So, among a few different methodologies that I use in my research, I
46 think the experimental one from what I hear now is the one that matches more or maybe
47 have more information regarding data management and because it's a massive
48 collection of information. What we do is we study a specific type of reactor and we
49 develop the experimental setups for this research but also the instruments, the
50 measurement tools that we will use to evaluate what is happening in in this type of units,
51 and so of course when we are focusing in the instruments, then there's a lot of tests and
52 so on. Not everything is leading into a result because it's a development, is like just for
53 the next step but when we have everything in place and we start researching on a
54 specific reactor shape or process, then it's about generating a lot of information with
55 this, with this measurement tools that we have.

56 **Interviewer:** Okay, very good. Let's think about a specific research project that maybe
57 you have in mind. If you look at the wheel and think about the data management

throughout the project before, during and after. Let's walk through the different stages and see how things went for you, what was important, what was working well and not so well. And we have a few recurring questions on the slides that you can refer to like what do you usually do with your data, and how does it work, what tools do you use, and what data do you store, and when, also this is associated to costs for you, and do you get or need any type of support, and then also if you encounter problems in that. Talk freely, if you want you can orient yourself a little bit on these questions. Let's start with the planning phase.

Participant: So, if I'm focusing on this experimental type of work... We do have discussions before we go to the lab, before the actual campaign starts, regarding the experimental metrics that we will work on and regarding what type of data will be collected and how this is going to be used to get our conclusions at the end. Everything thinking of how things will contribute to a final conclusion. Yeah, so this is how we plan. Regarding the data, depending on the analysis that we pretend to do with every type of measurement, we need to check, for instance, at what frequency we're going to measure. Sometimes that will make the difference between some information being useful at all or not. So, this is relevant. Also, for how long you log data because in order to be representative, you would need to cover also a length of time that is representative for the process. And that sometimes we have information or we have some idea about that and sometimes we lack this information. Because it's something completely new then we don't know what are the natural frequencies of the system that we are observing. So, we talk about, yes, for how long we measure, what frequency we measure, at what times during the experiment we gather data. This goes for planning, yeah.

Interviewer: And thinking about also the practical aspects of data, do you have a fixed plan that sort of always works for how to store this large amount of data? Or do you get in contact with, I don't know, IT department or somewhere?

Participant: No, definitely not. And this is, in fact, this is a problem because it grows like year after year, our capacity for measuring. And so, it grows the amount of data that

87 we collect. And we haven't formulated a strategy that accommodates this growth. It has
88 been quite simple. Normally there would be some space in the cloud and every person
89 that finishes a campaign will select the data that they want to save, and they will make a
90 folder and leave it there. But as the data grows, this is not suitable anymore. So, we
91 have started using just hard disks for saving data. Storing it for later.

92 **Interviewer:** Essentially, you're saying your group has internally found solutions that
93 always work for the project.

94 **Participant:** Yes, exactly. And even though we have had the idea, or we see the need
95 for some more structured manner for storing data, it hasn't been obvious for me what is
96 a good solution. I can like, I have talked to people internally at the division, and they
97 have some proposals but it's, yeah [hesitates] we're just filling space and yeah, we don't
98 have an actual strategy for future.

99 **Interviewer:** Yeah, good thank you. Okay, let's continue to the create/collect phase, the
100 second the gray one. How are you thinking about that?

101 **Participant:** Yeah, that for me would be the experimental campaign itself, yeah
102 [hesitates] it's all about producing relevant data. Yes, well, relevant is always a bit with
103 the question mark and so, of course, it is planned but there is still some uncertainty on
104 how useful things are going to be. Yeah, but this is the actual experiment collection of
105 data.

106 **Interviewer:** And do you sometimes encounter problems there? I mean, I think it is the
107 core or part of your research, right? But still, do you encounter sometimes problems that
108 are more issues that could be solved with support infrastructure?

109 **Participant:** Yeah, definitely. We do have people helping us, it's the manner in which
110 we work that when we develop instruments. We normally have a partner who is more
111 expert in the type of measurement we're attempting. So, we have some support. But in
112 fact, as the reading frequencies grow, we encounter problems. Either with the
113 acquisition systems that cannot support higher frequency or when you, during the
114 measurement, you might already process some of the information, right? We use

115 regular tools like LabVIEW for processing on-site. It's not really processing, but it's
116 manipulation of the data on-site. So, LabVIEW would normally work, but on the
117 hardware side, we sometimes have some problems. Some measurements become just
118 not possible, either at the frequency or for the length of time that we envision because
119 of just volume of the data.

120 **Interviewer:** Okay, so again, the size of the data is a big problem.

121 **Participant:** Yes, yes.

122 **Interviewer:** Okay, the next step, Analyze. Is that still something that is part of your
123 projects?

124 **Participant:** Yeah, it is all the time. The raw data is never the final result. It's, we always
125 need to process it. And we also develop the tools for the processing. And it's normally
126 part of the work, of course. It's what we have developed in the past is used in the next
127 project, and so on. But there is a continuous building on further tools... [hesitates] Yeah,
128 and I think in that part, we do like some capacity, in terms of, this will rely on the people
129 that is working in the project, whether they have enough time to invest in this and in
130 creating or improving the tools, but also sometimes in this maybe background and
131 understanding where this information is coming from like, what type of measurement it
132 was and what the specific characteristics it had. There we will lack some, some
133 knowledge. Yes.

134 **Interviewer:** Okay, I have a question regarding the data handling that maybe relates to
135 both the create collect and the analysis phase. Because you said sizes of the datasets
136 are increasing, do you usually use compression and decompression or are these sizes
137 that you can just handle as they are?

138 **Participant:** Sometimes it's not the typical procedure to compress data, but sometimes
139 we do.

140 **Interviewer:** Okay.

141 **Notetaker:** Okay, I have a question. Can you quantify the data size, the size?

142 **Participant:** Difficult... [hesitates] We are talking about (...) so what we do, megabyte
143 sizes we can handle and that means that we need to shorten our experiments or
144 logging times to that type of size. So, we could generate things like scale 10 or 100
145 more but then we could not handle them. We need to shorten the times and that, of
146 course, is different for different techniques, so sometimes it means we can only
147 measure for two minutes or sometimes we can measure for half an hour and this is still
148 okay. But this is what stops us from growing in measuring time.

149 **Interviewer:** Okay. So, let's talk about Making Available. I assume, usually it's not the
150 raw data that you, I mean you maybe need to share it in order to collaborate but then for
151 publication it is a subset. I guess that is how does that work for you?

152 **Participant:** Normally in what we deliver as a result, the information is already very
153 much post-processed. Meaning that out of hours and hours of measurements then you
154 collect maybe 10 points and so the result after all these processes, processing stages
155 then you come down to maybe one number or maybe, yeah, one correlation. And so,
156 the very big step goes through the analysis, then you go from those like massive files
157 into a very simple result.

158 **Interviewer:** And then is it not common in your field to publish the raw or cleaned raw
159 data such that others can reproduce?

160 **Participant:** No, no. What we normally do is that we would, our publication would go
161 around the very specific information that delivers the concept, delivers the main
162 conclusion. But we normally have broader experimental matrices and that information
163 would not go in the publication because it doesn't equally contribute to the conclusion or
164 it would be repetitive in a way. So that can be the wider experimental matrix with the
165 results can be made available as additional material in the publication. Yeah, but the
166 raw data, we typically don't make it available unless someone asks.

167 **Interviewer:** Okay, I'm not sure I fully understand the difference between this wider
168 matrix and the raw data.

169 **Participant:** Yeah, so if you have an experimental matrix, just a very typical, let's say,
170 10 times 10, so two variables alone. But we do this because we don't know exactly
171 where we're going to find, what we're looking for. And let's imagine already a four times
172 four would deliver the message. If you start just repeating the same type of plots, it will
173 not contribute to the conclusion in the paper. We normally prefer if we can summarize
174 this in a different way, then we put it in the paper. But if it's not necessary it will not go in
175 the paper so there are additional results. When I say, like, two hours measurement
176 result in the number then instead of publishing the hundred numbers, we publish only
177 20 numbers, something like this. To make the conclusion and the rest of the information
178 for other experimental conditions that would be provided as additional material. So, for
179 the for the actual phenomenon that you're investigating, there will be more information
180 but it's already post-processed.

181 **Interviewer:** Yeah, one question from [notetaker].

182 **Notetaker:** Yeah, I think you answered that with that last sentence that the (.) even the
183 stuff in the supplemental materials of a paper is already post-processed.

184 **Participant:** Yes, yes and very seldomly, we publish maybe just figures to explain how
185 the data comes out from the measurement. But this is just to exemplify what we do to
186 the data. But it's not(--) Normally is in the form of some plots. But we(--) Not really the
187 data, not the information that someone could process.

188 **Interviewer:** Okay, now we talked a lot about publications going one step back into
189 the... [hesitates] data sharing to collaborate that for you works via the cloud or, because
190 you said now you also have hard drives where you store the data. Does that mean
191 everyone comes in or if several people work on a project, they come in, they get the
192 data to their own computer and continue with that?

193 **Participant:** Yes, exactly. Whatever we can store in clouds, then we keep it like this,
194 and it would be made available for the collaborators. And otherwise, if it's larger
195 amounts, then we need to physically transfer the hard drive.

196 **Interviewer:** Okay, thank you. Data preservation (.) How does that work for you? Do
197 you preserve the raw data?

198 **Participant:** Yes, but without much of a system for it. [hesitates] Yeah, so if I want to
199 find data from someone else, then I will probably be able to find where the data is stored
200 but there is no system for like how this person stored this. Yeah, it's very down to the
201 researcher how they do it.

202 **Interviewer:** Okay. And about the archiving, have you been in contact with [university]
203 about this institute-wise archiving or is it all, like you say, locally?

204 **Participant:** Yeah, it's local. No, I haven't like dug farther into the, yeah, what possible
205 solutions could be there.

206 **Interviewer:** Okay, let's move on to the last phase: Reporting. Do you, I mean we have
207 the publication, do you in any other way report what the data looks like? I mean this
208 goes a little bit along with archiving, but what it looks like, where it is to be found and
209 things like that?

210 **Participant:** On the very surface, of course, for instance, when a student leaves and
211 they leave behind the data that they have collected, we would try to ask them to have
212 some system and communicate the system with the data that they leave, some
213 documentation that guides next person for the use of the data. But again, if you ask me
214 I would ideally have some sort of structure or form that I give to this person, to this
215 student and say please, do it after these guidelines. Instead, we simply ask for it but it
216 will be the person who decides how they do it. So, they do report, normally every person
217 that leaves, if it's a student, they would have some sort of report of what they leave
218 behind and there will be a meeting explaining things. But other than that no, no (.) In
219 some cases, in some cases when we have collaborations with industry and then we
220 need to report to the industrial partner, if it's, for instance, something that it was
221 measured at their site, then it depends. Sometimes they want to know more details
222 about the data, how it looked, and what they can do further with this data and so on.
223 Then we would have reporting between parties.

224 **Interviewer:** Okay. And connections to the outside of research world, do you do any
225 public science? I don't know, news articles, posts, podcasts, something like that, do you
226 do any of that in your research?

227 **Participant:** No.

228 **Interviewer:** Okay, good. And then one point that came to my mind, which we haven't
229 mentioned, is data management plans. Do you usually(--) Is that part of your planning
230 phase? And do you feel that they helped?

231 **Participant:** No, it's, again, it's very, very immature in that sense, the planning.
232 Because we are aware when we plant experiments, we are aware of, yeah, this
233 eventual need for some structure in the data and some structure for storing and so on,
234 or transfer of data. But no, it's very, very immature.

235 **Interviewer:** Okay. I understand. Is there anything(--) Please take a moment to read
236 through the notes that were taken and let us know if there's anything that you wanted to
237 add or that we should –

238 (...)

239 **Participant:** Yeah, I think it's well summarized. I would only add maybe one thing to the
240 analysis part, that sometimes we also lack capacity for processing. And we don't have
241 established procedures on how to do. We have in some of the projects used the
242 clusters to speed up the processing, but this is, in my opinion again, it's not optimized.
243 Because basically we just look for more capacity and let it run whatever this capacity is,
244 but it would be nicer if we have a better understanding on, for instance, how the data
245 could be prepared to actually process it faster instead of just like, yeah, let more
246 computers run our data.

247 **Interviewer:** If you say lack of capacity, that is maybe both the compute capacity, but it
248 is also a little bit of the knowledge of how to optimize.

249 **Participant:** Yes, yes.

250 **Interviewer:** Okay. Well, great thanks for that add-on. We are approaching the end of
251 the interview, and I would like you to now again go through each of these phases and
252 think about how you feel or how you felt in these phases. Are you happy with your
253 experience in them then we'll put a smiley face, is it neutral or are you more unhappy
254 and that really should reflect your experience. I'm often asked: should it be regarding
255 the support you get and it's not necessarily that because if you're happy, then well
256 either support has worked or you didn't need it but if you're unhappy then we know that.
257 No matter if there's support, maybe you didn't know about it. So, I mean there is an
258 issue even though it might exist so it's really an estimation of how you experience these
259 phases.

260 **Participant:** So, I must say that the last three stages, even though I am aware of the
261 importance, they do not represent for me any frustration if I don't have a good method
262 for the reporting or the preservation of data, it has less impact on how we progress. But
263 I think in... Indeed, like between collection and analysis I, this is more in my daily life, so
264 it gives me more frustration if I cannot do things properly there. Or I wish I could, I wish
265 to do something, and I don't have the tools and, yeah.

266 **Interviewer:** Okay, so does that mean, say for Create and Collect, we put a sad smiley.

267 **Participant:** Yeah.

268 **Interviewer:** And the same for the analysis phase.

269 **Participant:** Yes.

270 **Interviewer:** And then you said the last three you're aware, but it doesn't hurt you so
271 much, so should there be a neutral?

272 **Participant:** One neutral yes.

273 (...)

274 **Interviewer:** And then we have the Planning phase. Do you think about this one? And I
275 want to emphasize that we often tend to try to be nice and not put too many sad faces

276 but we are interested in your honest opinion. We're not personally affected, so please
277 be honest.

278 **Participant:** Let's see, so what I'm thinking at the moment is that I believe we have,
279 like, the discussions and the considerations we take prior experimentation. They are
280 good but again I, that could be done better if the next phases had better tools. So, I
281 guess the same type of dynamic could be used but it(--) The planning would be more
282 successful, if we had in the creation and analysis, if we had already some better tools.
283 So, it doesn't give me negative feelings, but it will be impacted so, or, it is impacted by
284 the other two.

285 **Interviewer:** Yeah, so do I understand neutral?

286 **Participant:** Yeah, I would say. Yes. And I also want to say that, of course, the creation
287 and collection of data and analysis, of course, it's a sad face. I agree. But it's also what
288 gives us our results and that makes us most of the time happy.

289 **Interviewer:** Yeah, I know. It's never just one thing.

290 [both laugh]

291 **Participant:** Yeah, no, but when it comes to data handling, yes, definitely it's a sad
292 face.

293 **Interviewer:** Okay, yeah, I understand. My last question: please describe your dream
294 scenario and it can be anything that comes to mind now thinking about this whole data
295 process. I mean it can even be small changes in some part of it what would be your
296 dream scenario.

297 **Participant:** Okay, well. Okay, very dream scenario I like, it would be perfect if we had
298 tools in place that help us structuring what we're going to collect and meaning that,
299 when the collection of data takes place, data has already a place to go. Yes, and that of
300 course means that we would design this structure beforehand. Yeah, and we would
301 have this place, this structure already where data goes. Also, from where data can be
302 just retrieved every time that we need something. So, in my vision it would be especially

303 nice that when you retrieve data, you can automatically get it from the stage or the
304 different stages it has had in the processing. And I'm thinking of this because normally
305 the processing has several stages and sometimes our capacity, abilities at the moment
306 take us to some specific stage but a couple of years after, we know that we could take
307 the data and give it one more turn and extract more information. But again, it's the
308 problem on how did we plan the collection and maybe that it was not exactly what we
309 wanted, maybe it was not correctly stored or...

310 **Interviewer:** Yeah, yeah, I understand that that will be very useful for you in the, like
311 long-term setting where, yes, research might continue a long time after it has started.

312 Great thanks a lot do you have any more questions?

313 **Participant:** I don't. Many thoughts but this is good. It has been a reflection for myself
314 [laughs].

315 **Interviewer:** Okay, good. Then I'll stop the recording.