

1 **Interviewer:** Maybe I can give just a short introduction about this project, the aim and
2 the goal. So, the aim of this project is to map how researchers work with the research
3 data today and what their needs are. Also, the goal is to develop the best possible
4 support for researchers, so that the researchers can focus on their research better.

5 The structure of the discussion from today, will actually focus on a simplified model. And
6 maybe we can see it in the second slide, if my colleague can show. So, this model
7 shows different phases of how to deal with the research data. And we have these
8 planning, data collection, analysis, accessibility, preservation, and reporting the data.
9 And specifically, we will focus on the capabilities, which is a broad concept that describe
10 process information, resources, competence, regulations, IT support, etc.

11 And yes, we also place special emphasis on investigating what e-infrastructure is
12 needed to manage research data well. E-infrastructure is usually not something you
13 think about it in your daily work until it is not available or does not work. But you need it
14 for your work as a researcher. Yes, maybe if we can go back to the first figure, we will
15 see here four levels that we will focus also, and maybe you can keep it in mind
16 throughout this interview. The first level about IT infrastructure, which is useful for data
17 storage, backup, network. In the second level, research data creation processing, it's
18 more like specific, and you can see here also research infrastructure listed below. Also,
19 we have more general category, which is the research data management in the level
20 three. Finally, level four, which is more information or guide to the researcher. So,
21 please keep this image in your mind throughout this interview. Yes, and I would like to
22 mention that later we will have the transcription of this recorded interview, if that's fine
23 for you?

24 **Participant:** Yes.

25 **Interviewer:** So, everything is fine? Okay. Yes! And we can start with the research
26 activity that's relevant for you. It could be a current or recent completed project. It is
27 usually easier to talk about a specific project, but if you come up with comments of a
28 more general kind, these are also very welcome.

29 **Participant:** Can I just ask you if you can go to presentation mode, because it's very
30 small on my screen? I have almost(--) It's almost difficult to read the text. I can have the
31 slide as big as possible? Yes, this is great okay yes sorry, I have a small screen in the
32 meeting room.

33 **Notetaker:** I think I cannot take notes on presentation mode.

34 **Participant:** Aha, okay.

35 **Interviewer:** Maybe you can just zoom it?

36 **Notetaker:** Yes, I can zoom on each part.

37 **Participant:** Thank you.

38 **Interviewer:** Yes, so maybe we can share again figure number two. So, we will focus
39 on this figure, maybe the slide number two first, just to show that this is our model that
40 we are following. And as you can see these(--) There is this life cycle of the data, maybe
41 the way to share it.

42 **Notetaker:** Okay.

43 **Interviewer:** (.) Yes. I think here. So, this is our model that we will follow in this
44 interview. We have these different phases of the research data. And the capabilities that
45 we will speak about in the gray boxes. So, feel free to speak about all these different
46 phases. And we will be happy of course if you tell us about these phases one by one.
47 And yes, maybe we can(--) If you don't have any other questions?

48 **Participant:** No.

49 **Interviewer:** Maybe we can start with asking, can you briefly tell us about your research
50 with focus on the data you collect or generate? Like maybe any project from your work

51 **Participant:** I am a postdoc now and I work with concrete samples, reinforced concrete.
52 But I work mostly with x-ray imaging so the type of data I look at they are tomographies,
53 generally volume of concrete samples. I look at corrosion development in concrete
54 samples and so it's generally rather large data. I have to use concrete samples that are

55 still in a reasonable size when it comes to comparison to reality. So, let's say, with a
56 volume that is 3-5 centimeters in width and maybe 20 centimeters in length. So, which
57 results in tomography volumes which are around 10 giga each, 10-15 giga depending
58 on the resolution so—

59 **Interviewer:** Maybe it's nice to say, maybe about the(--) This is from your last project
60 am I right or?

61 **Participant:** Yes! This is my last project, so that's my postdoc project I've been working
62 on this for the last two years.

63 **Interviewer:** Okay, so can we please like start with the first category which is the
64 Planning. How you do planning to collect this data, for your project? And you, can you
65 also [inaudible] in the box, some questions, for example, what kind of data how you start
66 to plan and if there is any collaboration for that.

67 **Participant:** Yes, so, it's experimental data. I follow the corrosion process of
68 reinforcements in concrete. The project consists in taking tomography data every more
69 or less three months of the samples and follow how the corrosion damage is spread
70 inside the concrete. The planning is, I have an experiment going on in the lab and every
71 three for months I take the samples and send them to [university] where they have a
72 tomograph that is capable enough that they can process my samples.

73 **Interviewer:** Okay, and is there any cost for only the planning phase?

74 **Participant:** Well, I do pay [university] for the acquisition of the data so, yes, the project
75 is paying for acquisition.

76 **Interviewer:** Okay, so it's included in the fund of the project and you get a collaboration
77 from [university].

78 **Participant:** Yes.

79 **Interviewer:** Okay, and do you face any kind of issues or problems in this specific
80 phase of planning?

81 **Participant:** Not really. I have to say the project is also in collaboration with
82 [Scandinavian university]. There is a parallel project going on in [Scandinavian
83 university]. I mean the planning is not a big problem. I mean when it comes to the
84 samples I have to ship them to [university] which had some issues, some delays in the
85 sending. But we also buy a tomograph in [university], so this is a probes issue that will
86 be solved in a matter of a year, more or less, so

87 **Interviewer:** Okay, thank you. So do you want to move to the next phase?

88 **Participant:** Yes.

89 **Interviewer:** Can you please tell us more about the Collect phase for your data?

90 **Participant:** The Collect phase (.) So, once the data are (.) Well, once the tomography
91 is obtained (.) the tomography data are sent to me from [university]. They have (.) I don't
92 remember what they use, I think it's OneDrive. So, we have a shared OneDrive account
93 where the data are stored. And then I download the data to my laptop and (.) Now I can
94 start talking about problems. So, this is very large data and I have time series so (.) So
95 what I try to do, what I'm doing, is that I correlate the time series and match them
96 together. Now I'm going into the Analyze phase. Collection is just I store them and back
97 them up. I don't have any space left on the OneDrive provided by [university]. So, I back
98 them up into separate hard drives.

99 **Interviewer:** Which kind of tools(--) Are you now speaking about analysis or you are
100 still(--)?

101 **Participant:** I was just on the collect, so the collecting ends up with the data on the
102 different hard drives.

103 **Interviewer:** Okay. If you don't have something to add and(--) Maybe you can see my
104 colleague what she's writing and of course you can change anything you think is not
105 correct. Yes, so if you don't have something else to share about the collect, maybe you
106 can tell us about how you do analyze your data, what kind of tools that you are using
107 and again you can tell us what kind of issues you face?

108 **Participant:** Yes, I use(--). Most of my image analysis is done in Python and I do that on
109 the cluster. I think the biggest issue is that I have a hard time doing like checking my
110 code and do local analysis. I have to reduce the quality of the data quite a lot to be able
111 to run something locally. And I get an output to test my codes. And when I work on full
112 scale images then the cluster works very well, but unfortunately it takes a long time to
113 download and upload data on the cluster. I have the process gets much, much slower
114 than it should be. And by the fact that I cannot really store all of my data, I mean I don't
115 want to store my data locally, but I have to have copies in hard drives. And I don't have,
116 I try to have, at least on OneDrive, I try to have the data I work on locally, but it's very,
117 very slow to download the data from OneDrive. I'm facing some time problems. I try to
118 update the RAM of my computer and my local laptop to make it at least faster or to be
119 able to run something locally, but I was told by the IT support that it's not something I'm
120 allowed to do.

121 **Interviewer:** And did you consider to use, for example, the HPC clusters at [university]?

122 **Participant:** Yeah, everything is done! I used a cluster at [university].

123 **Interviewer:** May I ask which cluster, like [name of cluster]?

124 **Participant:** Oh, it's [name of cluster]!

125 **Interviewer:** Okay.

126 **Participant:** It's [name of cluster]. I also have a small, I ask for a(--). How do you call it,
127 a small data(--). I have a small project, a small data project, something like that. So, I got
128 extra space on the cluster, with no backup, but at least it's big enough that I can run two
129 or three data at the same time and store the results.

130 **Interviewer:** Okay, so mainly the issue that you are facing, it's really time-consuming,
131 this process, too?

132 **Participant:** Yes, and maybe also storage. I still feel like it's a bit, not completely safe to
133 have all of my data on external hard drives.

134 **Interviewer:** Yes. But, you know, there is, like, a different location size, you can book.
135 There is a small, medium, and large.

136 **Participant:** Yes, I'm (.)

137 **Interviewer:** Did you get any support?

138 **Participant:** I have the small one. I got support from the cluster, but the small one
139 comes without backup. And we have, I think this is more like at my department level,
140 that we have decided to go for storage that is kind of limited for everybody in the
141 department. So, then(--). But we were thinking, we are looking into applying for large
142 storage as well.

143 **Interviewer:** Yes, this is what I meant, like I know now you have the small one, maybe
144 you can consider the larger storage. And you can get the support from [university] or
145 your department.

146 **Participant:** Yes, I had (--). I found it very helpful. I think the cluster is the place where I
147 have the least amount of problems. That works very well and it has been working well.

148 **Interviewer:** Okay, so you are using only Python to analyze or you have another tool?

149 **Participant:** Locally I use a software called ImageJ to look at the data and I also (--). I
150 never remember (.). Also I do also some rendering some 3D volumes and now I don't
151 remember (.). I will have to check, how is it called? (--). And RView for visualizing the 3D
152 volumes.

153 **Interviewer:** Nice. So, until now, this phase, there is no cost for that (.). since you have
154 the small storage. But if you consider the larger, maybe there will be cost?

155 **Participant:** Yes.

156 **Interviewer:** Okay, then if you feel like you said everything about the analysis, maybe
157 we can move to the next?

158 **Participant:** Yes.

159 **Interviewer:** How you make your data available? How you do this?

160 **Participant:** To other researchers? I mean, that's been a problem! The data is very
161 large. It's also uncommon in my field. It's rather new that we are using imaging, so it's
162 way too large for being something that can be in the storage offered by a journal. It is
163 often way too large for the database that we've been using for smaller data. We have
164 not really reached a conclusion on how we should do, to make the data available. I
165 worked in previous images, imaging program. I did my PhD at [university] and I also
166 worked with imaging there. And then we had an embargo period for the data so we
167 couldn't really share the data but now the embargo is finishing so we also have to think
168 about how we could share those data. I think we have been in contact with a Swedish
169 database. I don't really remember the name of it. It's (.) some Swedish research center
170 that offers a database, that needs to be paid for storage, that could be a solution. Still a
171 bit too large data. It would be possible to make available something, but when it comes
172 to making two terabytes of data available it's, the possibilities are very limited.

173 **Interviewer:** Okay. So, you are still thinking about this phase how we will make it
174 available in general, like as a summary for that?

175 **Interviewer:** Okay. So, preserve?

176 **Participant:** That's also a problem! It's on my hard disks, right now. Not the best way

177 **Interviewer:** Okay! So, you are using only local hard drives, to save the data?

178 **Participant:** I used also OneDrive drive but I ran out of space and I was told I cannot
179 have more. I used the one that I got but(--). And I saved those, the data I'm working on,
180 on OneDrive but(--). Yes, that's not enough to save all the data. Preserve is also(--). I
181 mean this, the OneDrive is linked to my account, so when I finish as a postdoc then the
182 data should go to somebody else, and that's a bit, like, I mean they will(--). On the short
183 time (.) where they will stay, in the hard drives. But yes, (.) we're debating on what to
184 do, to have still the data. This are also rather, it's rather new for us as a field. And we
185 think the data could still be worth to have around, and maybe we have, we're going to

186 acquire better techniques, or we could do other things with the data. So, we think it's
187 important, at least for the next like five six years, that we have them available.

188 **Interviewer:** Can you just explain, like, are you saving it as a raw data or you document
189 after the analysis and...?

190 **Participant:** I saved raw data and then I also save the finally, the final outcome of the
191 process. I mean I try not to save anything in between or try not to save anything that
192 could be easily obtained by running a code, so that it diminishes the size of the data.

193 **Interviewer:** Okay, but you are documenting this process in between? To help the other
194 researchers?

195 **Participant:** Yes.

196 **Interviewer:** Okay, nice. Maybe we can go to the Report, reporting this data. I think it's
197 (.)

198 **Participant:** So, report (.) Do you mean(--)) I think I need an explanation for (.)

199 **Interviewer:** Publish it (.) like, make the other researchers use it or, yes, what kind of
200 like, are you publish these data in articles or?

201 **Participant:** Yes, so now we are, up to now, have not published anything but one of the
202 articles is submitted. And second is in the writing process. Journal articles are what we
203 are planning to use to report the data.

204 **Interviewer:** Okay, so the articles. Okay maybe now, if you don't have anything to add
205 for these phases, maybe you can look at what my colleague wrote there. If you would
206 like to add something?

207 **Participant:** I have been reading a bit, so it looks correct.

208 **Interviewer:** Okay, great! I think, my colleague (.) has three phases, I don't know, I
209 don't see(--)) There's three phases in the bottom. We want you to add one of these three
210 phases, to each phase of the research data. To explain if you are happy about this part
211 of research data phases like if you are satisfied, neutral, or sad about it. And we request

212 you to be honest about that, like for example are you happy with the planning with the
213 creation, collect the data, analyze, and so on. So maybe we can pick one face for each?

214 **Participant:** Yes, so the Plan and the Collect part, I don't have anything I am(--) I like it!
215 I'm happy with it.

216 **Interviewer:** Okay, then maybe we can pick this to the Planning. What about the
217 Collection also, okay?

218 **Participant:** Yes! (.) The Analyze (.) there have (.) things I'm not happy with. (.) I am
219 happy with the cluster. The cluster is very powerful, it's really nice for analyzing! That
220 makes me(--) That I'm happy with. But I am not happy with the hardware limitations, and
221 the storage limitations. So, I guess I'm neutral!

222 **Interviewer:** Okay, so with the analysis you're neutral. Make it available, I think it's from
223 what I understood, it's the same.

224 **Participant:** I think Make Available, I'm unhappy because we haven't really found any
225 good solution. And the same goes to preserve. I'm also unhappy with preserve. Journal
226 articles are okay. I guess I'm happy with my journal articles.

227 **Interviewer:** Okay. What do you think about, make you more happy with the these two
228 phases, like Make Available and Preserve?

229 **Participant:** I think it's (.) it would be, well there should be a place where, it doesn't
230 have to be unlimited space, but that is given to old projects to, well to projects that
231 where they ended, you can wrap it up and save the data, all the things that are going to
232 be used in the future. But there is no(--) Nowadays there is no clear guidelines on what
233 to do once you are done, with what do you do with your results, and what to do with
234 your raw data. And I am happy with the storage possibility. It's, I think it's quite difficult
235 to be always limited by the space available. And I understand it's needed, one cannot
236 occupy a limited amount of space. But I don't(--) Compared to some other kinds of
237 research, my data are a bit large but they're not incredibly large. I think it should not be
238 just painful to move them around. I think the hard disk is not a good solution. It's a very
239 viable, if there's a fire in the office then everything is gone.

240 **Interviewer:** Yes, but do you think like you took some steps in order to solve this issue,
241 like did you just do something to solve this for the future?

242 **Participant:** No (.) I mean I talked with the IT, but I didn't really find many solutions. I
243 tried to find (--) I took some step for myself, like the small project on the cluster is
244 helping me with decrease in the time for analysis because I can still store things there.
245 But (.) I have tried to obtain larger cloud space, but I was denied all the time.

246 **Interviewer:** Okay. So, what do you think, what kind of(--) Do you think the
247 collaboration with outside your university can help in this issue or? You just contacted
248 the IT from [university] and you waited for the solution. Or do you think you want to take
249 bigger steps in this order?

250 **Participant:** I mean I don't think it should be on the collaborators to help with this. It's
251 for our research group that we want to keep the data. I mean we also should make(--)
252 Be made available to other researchers, but the problem is mostly internal.

253 **Interviewer:** Yes, and as you mentioned in the beginning, like if you increase the
254 storage, I think in the clusters that might also help. Are the department willing to pay for
255 that or, I don't know, how you deal with the cost if you want to apply for larger storage?

256 **Participant:** We have it, we have a talk next week, so trying to increase the space in
257 the cluster. But it's (.) it's a bit (--) And of course we could pay ourselves, but this is all,
258 we are still going in the solution for today, while I'm working on the data. But I think the
259 biggest problem is that there's no solution for tomorrow. Because I don't think we can(--)
260 I am happy to have my data on the cluster and for the working time it's enough with the
261 small space. But I cannot leave the data on the cluster for five years! So it's(--) The
262 biggest problem is that this data should be somewhere in three, four years, available.
263 It's, yes, that won't(--) That is not something I can solve with the larger space on the
264 cluster.

265 **Interviewer:** Okay. So actually, we wanted to have a complete description about the
266 dream scenario, like now you explained a bit about making available and preserve. Do

267 you think you have other things that you would like to have, like a perfect scenario for
268 dealing with your research data?

269 **Participant:** Yes, okay. So, a perfect scenario. I think one thing that me and some of
270 my colleagues have had problems with, is the limitation on the laptop available. So we,
271 it would be very good to be able to purchase it. I mean to have as a standard purchase,
272 a laptop that are more powerful, or to be able to change RAM, for example. So that's(--)
273 And especially to be able to run Linux, because we are limited to Windows right now. At
274 least we tried to have a [university managed] laptop with Linux but we were also denied
275 on that.

276 **Interviewer:** Okay. So, this is again back to the hardware and the resources that used
277 for the storage and?

278 **Participant:** This is, yes, this is on the analyze part, so that will make my analyzing
279 process faster. To have a(--) To be able to at least do local checks on my data or my
280 laptop, before running on the cluster. Because the downloading and uploading from the
281 cluster is a bit of a long process.

282 **Interviewer:** Yes. So, coming to the time-consuming that you mentioned, now this is
283 one of the solutions. Do you think there is more steps that you can take to make your
284 analyzers faster? For example, changing the tools using, for example, machine
285 learning. Is that helpful?

286 **Participant:** [sighs] Not (.) I mean (.) My impression is that(--) Tried to run a couple of
287 machine learning algorithms, and that were not fast! [laughs]

288 **Interviewer:** Okay.

289 **Participant:** I think that we still have to go to the cluster and, more or less the same
290 problem. I think, I mean, I think the way I run, I kind of optimize the way I run my
291 analysis by mostly using only Python, which is quite, I mean, there are softwares
292 available to look at images and extract data, but most of them are graphical. So, they
293 are weighted, they will just crash on the local computer. I kind of optimize the way I run

294 everything for the cluster. I think that's the fastest way. I mean, I also like how I do it. I
295 think other softwares would be worse in terms of time.

296 **Interviewer:** Okay, you mentioned that you are using Python and ImageJ. Are you
297 working until the pixel levels like, that's why it's time-consuming?

298 **Participant:** Yes, because I mostly work with (--) well yes, at the end is very very big
299 matrices so that's (--) I work mostly with the segmentations, and trying to extract data
300 out of the pixels.

301 **Interviewer:** OK, great! Thank you for the information. Maybe you can just take like a
302 final look at what my colleague wrote. Maybe we forgot to write something, tools or
303 something related to the analysis or any part of this.

304 **Participant:** Yes, I think it looks correct!

305 **Interviewer:** So, we are approaching to the end of this interview. Thank you very much
306 for your time and agreeing to give us this information. This definitely will help us to
307 analyze better and do our project better.

308 **Participant:** Thank you. Sorry if it sounds mostly like complaining but! [laughs]

309 **Interviewer:** No, we are happy that you shared honestly about everything. Thank you,
310 then, have a nice day.

311 **Participant:** Okay, perfect, I'm done, it's done. Okay, thank you. I hope it helps. Bye
312 bye.