

1 **Interviewer:** Okay. So, now we're recording and we can see the presentation. We're
2 good to go, I think. Alright. The interview is going to be sort of(--) We're going to try to
3 keep in mind these different aspects. You know, there's the IT infrastructure. There is
4 the work being done during the project in research data creation and processing. You
5 might be using some infrastructure there. There's the research data management tools
6 and infrastructure. And then finally, there's the support services. Like SND or like
7 [university research data support], for instance.

8 Alright. We can go to the lifecycle slide. And this is the lifecycle we're going to be using.
9 It's part of a tool called *Vintergatan*, which is sort of like organization mapping tool or
10 methodology, you could call it. Well, I think you're quite familiar with this by now, so I
11 think we can go directly to the interview and start with the planning phase. So the
12 questions will be recurring and, you know, of course, they're not always relevant. What
13 do you do with data during each phase? Maybe we're talking more about the planning
14 tools than the tools for data here.

15 **Participant:** Yes. In all these projects, I use the [names a data management plan (DMP)
16 tool].

17 **Interviewer:** Okay.

18 **Participant:** I also, in most projects, I already have some data, but I tend to collect, but
19 that's the next box. In terms of the planning, I would say, identifying sources and then
20 doing the data management plan, of course checking and deciding depending on the
21 project, what kind of sources, also based on their licenses, I should use. I do a bit of
22 research at the beginning, identifying, choosing, sometimes not deciding definitively, but
23 choosing valid or valuable sources. And then once the project starts, I do the data
24 management plan at the beginning, but there already kind of with a lot of information
25 and what is it going to be done what other sources where I store it etcetera, etcetera.

26 **Interviewer:** Yes, okay. You mentioned that you use the [DMP tool]?

27 **Participant:** Yes, yes, I use that from the(--) Go to the internet and start that and I've
28 used it in a few projects actually there's one that started recently that I still have to do.

29 **Interviewer:** Okay.

30 **Participant:** And that's sort of a (.) it's just that's the norm for me because it's helpful to
31 go through.

32 **Interviewer:** Okay, do you, do you encounter any issues here or any?

33 **Participant:** No, no. I just find it useful also because I noticed that when people on the
34 other end actually receive prompts, when things are added and edited, we eventually
35 get questions. And eventually, there's also support in a way.

36 **Interviewer:** Alright, yeah. Through the [DMP tool]?

37 **Participant:** Yes. That's a request as more information or to think a bit further. But yes,
38 identifying sources depending on the project, which ones I should use. There's many
39 criteria, licenses and openness if it's an open project for(--) Where I intend to make
40 everything open at the end, that's an important criteria if it's something where there's no
41 intention of that then I can choose other sources, etcetera.

42 **Interviewer:** Yeah, alright.

43 **Participant:** But it has a big impact like what(--) Because I reflect from the beginning, I
44 would not want to do all the work and then come to the end and realize "oh my god I
45 had planned to publish this online and now I realized that I've used the data that does
46 not allow it to be published on a web map", for example, that would be a disaster, so I'm
47 quite conscious and quite careful from the beginning.

48 **Interviewer:** Okay, that's great. In the interest of time I think we can move on to the
49 creating and collecting phase. And I wonder there again what...? [hesitates]

50 **Participant:** Here, from my research, we usually collect rather than create. We either
51 get access to open data, we purchase data or access data that is available for research,
52 access data through colleagues that are partners in the project that have the license.
53 Eventually, we might use secondary data that the partner has collected. That's the case
54 in one of the projects. We asked colleagues that we know that had created a dataset if

55 they could share and we can collaborate in the future. But we are very lazy. We rarely
56 create our own data [laughs]. We go through and really collect first.

57 **Interviewer:** That's effective.

58 **Participant:** Yeah [hesitates] of course what we do in the day then we work a lot with it,
59 in terms of processing, and I would say here it's not really(--) it's before analysis so to
60 some extent we create data by cleaning, filtering, adding attributes and augmenting. We
61 do a lot of pre-processing augmentation, linking different datasets. So, all these steps in
62 a way is that I would consider it to be a degree of creation because what comes out in
63 the end is much more advanced than(--) Or more relevant for the project than what we
64 had as raw data. In our databases we always keep(--) I, that's also a(--) I say that's how
65 we do things, we always have a schema for raw data, where we keep the raw things
66 exactly as we receive them, we don't touch that and then we have another for pre-
67 processed data or processed data as we call it where we just have the data that we are
68 actually going to use and it has gone through all kinds of transformations and this is
69 prior to analyzing.

70 **Interviewer:** Okay, that's good and do you use any specific tool sets for this, or?

71 **Participant:** My favorite is [hesitates] well for collecting of course there's a bit of
72 searching, so online with a browser and downloading. I use a lot of geodata so all kinds
73 of geodata portals and open geodata portals for downloading. In some cases, we get
74 data via APIs or other services that are automated. So that is also (--) we prefer
75 automated sources that don't involve us clicking on the interface. We just write a line of
76 code to retrieve the data that we need, that's the preference. Tools that we use for that
77 so it could be Jupyter Notebooks, could be Arc. I personally (--) it could be shell scripts
78 just to send requests from the, like to an Amazon bucket service to download some data
79 but also then the pre-processing my preferred place to do it is in the PostgreSQL
80 database with SQL. That's where we store the raw data, usually.

81 **Interviewer:** And you create one database per project, I guess?

82 **Participant:** Yes, I have got one big server where I have lots of databases for the
83 different projects, and every time I start a project, I create a database, I create a raw

84 schema. I load the raw data, I create a processed schema where I then put the results,
85 and then the rest depends on the project for the analysis and other operations and other
86 results.

87 **Interviewer:** Let's see, do you use any support here?

88 **Participant:** Yes, in terms of thinking back at the first slide, we use The PostgreSQL
89 server is installed on a server from [IT support]. I don't use any of the national services
90 because I'm all for compute or for storage, but I don't know if they(--) But it's storage
91 tied to compute. If I don't intend to use compute, then I don't use their storage. So, it's
92 essentially [IT support] server that I keep increasing and paying more for as I get more
93 projects. That's the server, and the rest is, yeah, no, the rest is, essentially, that's the
94 key thing. That's a really key server with the database.

95 **Interviewer:** And in terms of costs here, you mentioned that you have, you pay for this.
96 Where does that come from?

97 **Participant:** In the first few years, it came from my own research budget, which I had
98 initially for four years. Now it comes from the division. And I put it on the division, not on
99 individual projects, because the server is widely used. It's not only me, so I use for my
100 research projects. I made it accessible to other projects from other people in the division.
101 I even use it for geodata that we organize and process for teaching purposes. So there
102 are various databases there that go beyond my research projects.

103 **Interviewer:** And okay, we can take the next step then to Analysis. So, again, how does
104 the analysis process work for you and the tools that you use?

105 **Participant:** Yes. For analysis, we usually just use local computation on our local
106 machines, on our laptops or desktops, using Python or R or SQL on the server. When
107 something becomes a bit heavy(--) Essentially that also comes down to the researcher.
108 Some people love R, some people love Python. I personally love all of them, but I'm
109 very familiar with SQL. So, I often find SQL is always faster unless there is a complex
110 algorithm. A complex algorithm, then I go for R or Python. But generally, I manage to do
111 everything faster if I just do SQL. And faster because I'm talking about scaling up, doing
112 bigger, more things, rather than a small sample.

113 **Interviewer:** I see.

114 **Participant:** Yeah. And that SQL also has the advantage of running on the server. If
115 we're doing a long process, which is another issue, and we have that, processes that
116 take several hours or could even take more than a day, then obviously the server is
117 preferred because we can just disconnect, go home, and leave it running. With laptops,
118 if people are using Python or R, it's always an issue. We've had projects where we run
119 into that, and then people have to manage when they start, they have to stay a bit
120 longer, they have to leave the computer running in a cupboard and stuff like that. It's not
121 ideal. We are also setting up the server to do some of the analysis. So I've installed a
122 Python environment where we can have some packages. We still haven't really used it
123 for that, but we will, essentially, because of these issues.

124 **Interviewer:** And do you do all this by yourself, or do you get any support from IT?

125 **Participant:** I get support from IT on the server itself. Meaning some management
126 tasks, if it needs to reboot or if it needs Ubuntu upgrade, stuff like that. If I have
127 questions, actually, I have a question to send them about user access and stuff like that.

128 **Interviewer:** Okay.

129 **Participant:** But rather on the more technical side of the server, maintenance, if I need
130 to increase memory or disk, stuff like that. The rest, management of the server itself, I
131 do it all. They do the backups. So, when, that's the management of the server, I ask
132 them to do daily(--) There's daily images of the disk being stored but that's it. Analysis,
133 yes, that's how we do it.

134 **Interviewer:** You have, like, the server with the raw data, but do you also use it as a
135 working storage?

136 **Participant:** Yes, good point. It is the production server, let's put it that way. I keep the
137 raw data there just if we need to go back some steps so it's not for archival. For archival
138 I have hard disks. I have several duplicates of the data in hard disks in a safe, a
139 physical safe in my office. Actually, that's an important aspect. Also, because some of
140 the data(--) So, there's a management of it in the server itself, some of the databases,

141 there's different users with different access levels. And not everyone has access to all
142 the projects. And some that use data that is more sensitive or at least has personal
143 information then only the project participants are allowed to access. And even then, the
144 raw data(--) They usually do not even use the raw data. I'm the only one who
145 reprocesses the raw data in those cases. If the raw data, in those cases, the raw data is
146 there only for a short period because also it's quite big so I only leave the pre-processed
147 data in the server.

148 **Interviewer:** Right, right, and how do you experience management of data here in
149 these phases? It sounds like you have some load on you then too.

150 **Participant:** Yes, there is some load but it's also so specific to every project, if I mean
151 of course there are these things "we always do it like that" but it's us that decided to do
152 like that, there is no set way of doing it and I would not want to adopt someone else's
153 way of doing it. I kind of, I mean, we're always happy to learn and I do look at other
154 person's workflows and examples and GitHub stuff, so I'm always happy to learn.
155 There's even new tools for data processing that I start hearing about that I wonder
156 should we use this instead and then we sometimes test new things. But we are very
157 much hands-on and take care of things. It's very data heavy work that we do in terms of
158 analysis and so we need also from researcher responsibility we need to be fully in
159 control and aware of what's happening to the data throughout its life cycle when it
160 comes to the end not knowing how things are done or why would be putting the
161 research into question.

162 **Interviewer:** Yeah, I understand that, okay. Is there anything you would like to add to
163 this analysis phase?

164 **Participant:** Are you putting in there documentation, I mean we keep, because
165 everything is code based, there is code of every step along the way that we keep in
166 GitHub repositories. Eventually they become public when we have a working version of
167 all the messy steps that we do along the way, but within the team we can always go
168 back. We have the notebooks, so we know what was done. There's other ways also.

169 We have regular meetings, so we have ongoing PowerPoints where we collect
170 information as we progress, etcetera.

171 **Interviewer:** Interesting. Right. So, yeah, I think that's the analysis bit done. Making
172 available. And I guess this is sort of like sharing and publishing.

173 **Participant:** So, Making Available, there's various levels. Of course, the data that is the
174 output, the results, we try to make it available open whenever possible. We have
175 recently published something in Zenodo.

176 **Interviewer:** Right.

177 **Participant:** And, yes, I would say that we go for Zenodo just because it's more global.
178 We could put it in other places other type, other intermediary pre-processed data. We
179 have now in our department created this catalogue or repository for data sharing and
180 what we are now in the autumn doing is some of the data that we pre-process, we are
181 sharing it within the department at least. Because it just saves people a lot of time going
182 from a messy raw format to something that is more usable. If 20 researchers in the
183 department have to do the same and spend the same amount of hours doing that work
184 it's a waste of time and so we kind of try to share some of these datasets that we think
185 are of general interest, even just mentioning that we have them and if people want more,
186 they can contact us. And then one principle that I follow, actually, maybe this is
187 interesting, is that when I share, I always make it available openly if the data allows. If
188 someone sends me an email: "oh, can I use this?", for me, that's a trigger to make it
189 open. Meaning, okay, if someone sees value in this, so maybe others will. I'm not the
190 kind of person to send privately by email, hoping for a citation or an acknowledgement
191 only. Because it's almost the same effort, in a way, because I have to prepare it to make
192 sure that the data is there, it's clear, readable, etcetera. And with less impact. So when
193 someone says, "oh, can you please share?" Then I say, sure. And then I make it open
194 and say, "oh, and by the way, here's the link". And I send an open link.

195 **Interviewer:** Yeah, that sounds reasonable. You're still doing the same amount of work
196 for preparation, more or less.

197 **Participant:** Yes, and in that sense, also avoids the data being in a drawer not being
198 used. And if someone sees value, maybe others will see value. And rather than giving a
199 privilege to just one person to have the data, so removing that sort of artificial
200 advantage, let's put it that way, of the person who dares to ask me.

201 **Interviewer:** And let's see here, sharing. So yeah, you're sharing, you have your data
202 stored locally, and then you also make it available via Zenodo which is shared, so that is
203 the storage bit, the cost. Well, cost, I guess, is the cost for preparing?

204 **Participant:** Yes, it comes within the project if we manage to do it within the project
205 budget.

206 **Interviewer:** And when it comes to support, then you're not really getting or looking for
207 any support?

208 **Participant:** No, for Zenodo, no. And then, of course, some publications are more
209 demanding. Recently, we worked on a on a data paper and that was for a biology
210 dataset that was extremely demanding, the standard that they expected. The metadata
211 that the journal required was quite demanding so it actually was quite a bit of work to
212 put it in the shape and all the information that they required so it made the whole
213 publication process much, much longer but it was a data specific publication.

214 **Interviewer:** I see yeah. Let's see here. Alright, let's keep going to Preservation then.
215 You mentioned already a little bit that you have an archive or repository going on there
216 so can you tell us a bit about preservation and what tools you use?

217 **Participant:** Okay, so at the moment, I would say the project databases, they are still in
218 the server. And eventually, when a project finishes, I might make a dump. I don't have
219 so many that are what I consider closed to make that dump. And I keep also a backup
220 of that database dump in a disk. And if I find that the data and the results can be useful
221 in the near future, I keep it in the server. So that's for the results. The raw data, as I said,
222 irrespective of the preprocessing and the project, I keep it in disks. And this new catalog
223 archive that we have for sharing, in a way, also preserves and shares the data within
224 the department. I call it catalog slash archive or slash repository because it's not meant
225 to be very big in terms of storage. And so, in many cases, there will be an entry just to

226 signal this is available, but there is no download. There might be a sample, so people
227 know this is available; this is how it looks like. But the data is nevertheless stored either
228 in the server, in this database, or in the disk. I don't store zip files or stuff in the server.
229 The server is for the database. So it's not a store. I use it for processing. I don't use it
230 for archiving. That would be something that I could use instead of the hard drives, but it
231 costs money. I have some stuff that is non-sensitive, so some smaller non-sensitive raw
232 data downloads and stuff. I also create teams in Teams, therefore there's a file area. So,
233 for every project, there's a, if not a team, at least a channel within the team for the new
234 project. And within that file area, we might keep the meeting notes, some secondary
235 information of the project, and eventually some data. So some raw data that might also
236 be in the disks, but I keep it close to the project to be easier to find.

237 **Interviewer:** You mentioned initially that first you have a project database in the server,
238 and then you dump it. Do you dump it to a hard drive, or do you dump it?

239 **Participant:** Yes, so I dump in the server and then extract that and clean it essentially
240 the server I need to keep the size under control because it's not for archival like I said
241 it's for processing. A new project comes in, we're going to do analysis, we sometimes
242 run out of space and so that's why I bought several hard disks where I keep the archival
243 data.

244 **Interviewer:** Yeah, it sounds good and you mentioned already the costs here that this is
245 departmentally costed part.

246 **Participant:** Yeah, the server, yes, and of course the hard disk, it's small costs the hard
247 disk. I buy from the division budget as well if I could put on some project it depends on
248 the project budget as well. But the hard disk is the archival I would say; it's a small cost
249 at least at the moment. And the safe, I didn't pay for it [chuckles]. It was already in the
250 office when I moved in. I don't know how much it costs, but it's been very used. Some
251 different researchers and different projects have been using this physical safe.

252 **Interviewer:** That's good. Okay. Let's see. Anything else here? Did you encounter any
253 problems or issues in this phase?

254 **Participant:** The preservation?

255 **Interviewer:** Yeah.

256 **Participant:** I mean I don't think it's ideal to(--) Because I have multiple disks, I have to
257 manage everything myself, I've wasted a lot of time transferring between disks
258 increasing to bigger disks, finding the best format. I found that for example because I'm
259 using a Mac if I format the disk in a specific format, it's so much faster to load and to
260 save data, so I spent a lot of time. The main thing is I spend a lot of time to keep track
261 and to make sure there's the two backups and that they are correct and complete
262 etcetera and so yeah, I've spent a lot of times, but that's the main ones.

263 **Interviewer:** Okay, finally Reporting, this reporting one isn't so clear, but I guess this is
264 reporting back on the data and on how you use the data, is there even a mechanism for
265 that at [university]?

266 **Participant:** No, but I mean this new data catalogue in a way is also intended of like a
267 moment to announce what has been, and publish what has been, data that has been
268 the output, the result of a given project and we are even discussing that it will be
269 mandatory for PhD's and postdocs to put the data there once they finish so it's like they
270 don't get a diploma until they check out, the same way that they have to return a
271 computer they have to return a phone they have to put the results of their work in a
272 place that is available.

273 **Interviewer:** And when you say checkout, is there some sort of like git system involved
274 there or is it just?

275 **Participant:** No no no, by checkout I just mean there's an excel list saying this is what
276 you need to do that the personal administrators have and this will be an item added that
277 they need to(--) The data, now the thing is the only dilemma we don't [hesitate] we are
278 not sure yet how it's done because partly the cataloging can be done, so these are the
279 datasets that they've used and the datasets that they've produced that they can catalog
280 in there. But as I said it's not a big storage space so it's not a repository for with endless
281 capacity. Therefore, one still has to use hard disks but cataloging it makes it more
282 formal and then we also are aware that we need a disk with the data that it's in the
283 catalog and we need to develop a procedure for that. I know that in some divisions they

284 have shelves with a lot of hard disks from data, with data from the people that leave
285 from the different projects. But that's other divisions. Like, yeah, they have a [hesitates].

286 **Interviewer:** Data library?

287 **Participant:** Like a library with disks of the data resulting from projects.

288 **Interviewer:** Okay. Yeah, alright. And you've mentioned here about the problems even
289 of science then and of keeping track of all the data. We'll move it on then. So, two things
290 left. The first is dream scenario, and like a perfect world, what would be different from
291 the state?

292 **Participant:** Okay. In a dream scenario, I would have, instead of my safe, some
293 archival server where I could store all the data that is used as input, and also all the
294 data that is output. So, no hard disks. That would be one. Of course, that server would
295 need to be backed up as well. I would be reassured that even if something happens,
296 there would be a backup. If I by mistake delete a folder, which has happened, for
297 example, by mistaking this manager. Actually, I forgot to mention in one of these, my
298 processes of managing and moving folders of raw data between disks. Once I deleted
299 one before it had been moved, I had to contact the supplier again. And luckily, they
300 were kind to say, yes, we make it available again for you for a week so you can
301 download it. But they were kind because the data should not be made because I
302 created work for them. I was lucky. I was lucky once.

303 But indeed, that would be an important feature to have a backup. So that is an important
304 one, and I think the one that people think most about, this archival, long-term archival of
305 the input and the outputs. The rest, to be honest, the way we use the server [hesitates] I
306 keep having to increase the memory. So ideally it would be one with fluid capacity or
307 storage capacity that I could use as much as I would like. Maybe then I would abuse
308 [chuckles]. I don't know. So and this is for the processing server where we store the
309 working files. I mean, the archival one would have to be quite big anyway. The
310 processing one would not have to be so big, but would have to be more flexible
311 because sometimes it requires, yeah, it requires just to add and purchase more disks to

312 be added. I don't know how it, in some cases, it takes a few days because they want to
313 change the disks instead of just adding to the stack.

314 **Interviewer:** And I guess with the processing server, then an argument for having a
315 processing server rather than using like computing resources available elsewhere is
316 that you have the processing server set up as you like it?

317 **Participant:** Yeah, the thing is the computing services, I've been to some sessions and
318 they are good for very specific tasks and specific code. But not for everything. So they
319 are, in this case, for example, I want to process in SQL in a PostgreSQL database, and
320 that's not something that they have or want to install there. So that's one of the reasons.
321 Also, I probably would benefit from more cores or CPUs in the server. I'm paying for
322 very little. I could benefit from more. But this sort of database does not benefit from
323 GPUs and stuff like that. So, it would not, this kind of calculations do not benefit from
324 the GPU type processing. I think the compute servers for specific types of computation,
325 that it's not necessarily the one that we do.

326 **Interviewer:** Okay.

327 **Participant:** I'm always thinking, whenever I have a project with big computation tasks,
328 so I was thinking, "ah, is this now the time to use the compute servers?". I have to say,
329 there the barrier has been, I mean, the request, the setting up, I've done it once. For
330 small ones, it's fine. But then I realize that you have to renew every year, and that
331 becomes annoying. and I usually(-- I would love to keep it for multiple years because
332 projects run over 3-4 years and even after the project finishes then I would not have a
333 place where to put it, so I don't want to create something on a specific server where I
334 don't have then equivalent storage for archiving it so for that reason I also not going in
335 that direction. So, yeah.

336 **Interviewer:** Okay, one last thing. We're supposed to sort of like go through the cycle
337 and in terms of satisfaction like how well does it work? Does the three smileys back
338 there or smile neutral and frown, place one for each phase. How would that look like
339 today?

340 **Participant:** Okay, the planning phase, I'm happy with the planning phase, to be honest.
341 The create/collect phase, ideally, I would have already access to [hesitates] so it's like
342 so-so in the middle. And I would love to have a better server, of course, but also to store
343 the storage and the management. And this is where a lot of this hard disk backup work
344 happens, is when I start getting the raw data, then putting in the server, pre-processing,
345 and I have to manage all that. Actually, maybe that is an unhappy phase there. It's a
346 more cumbersome phase. When it's done, then, of course, the preserve, then it's a so-
347 so phase. So, the unhappy is in the second stage, the create-collect, sorry. I have
348 changed my mind [chuckles].

349 **Notetaker:** Sorry, can you repeat, please, for the last?

350 **Participant:** Unhappy on the second stage, create/collect. Because that's where most
351 of the hardest backups and raw data versus process data are happening, and where
352 even mistakes happen. Then, in the analyze phase, because for the very specific type
353 of analysis we do, that's fine. I would like the server to be better, faster, etcetera. I will
354 put an OK, the middle one. I'm okay having to manage it because everyone does
355 custom work. I prefer to do it rather than be restricted to whatever compute server
356 environment is given to me, so I prefer to be in control than to be restricted even if it
357 means more work. Making available, I would say I'm happy there is it's work but we do it
358 for ourselves. I accept that it is necessary and beneficial. The preserve I would say it's
359 [hesitates] I would put the(--) I'm not happy, because ideally we don't have archive
360 essentially, so I'm not happy. I'd say I'm not happy, yes there is less work now at this
361 stage, I just skipped the drives that I had at the beginning, but I don't think that's the
362 best solution. Ideally, I would move from the database server to another server without
363 involving physical disks. And then in terms of reporting, I'm happy, I've of course, been
364 quite involved in this new catalog, which I think will help a lot. And in any case, we
365 publish in GitHub, we publish in Zenodo. And so, we do it anyway, in other means. So
366 it's part of our work to report.

367 **Interviewer:** Okay, well we have seven minutes left so I will say thanks a lot for your
368 time and I'm sure we'll cross paths soon again and I'll let you go to the next meeting
369 now.

370 **Participant:** Thank you, thanks a lot.