

1 **Notetaker:** I'll be taking some notes on screen, so you'll be able to see what I'm sort of
2 writing.

3 **Interviewer:** So if we start at, I think maybe it's good to take a typical research project
4 maybe one you recently finished for example, if you can take that into mind. If you then
5 look at the planning phase then there's some questions like what do you do, how do you
6 plan, the (--) what you do with data do you have a data management tool etc. How does
7 this work in your (...)

8 **Participant:** Well as, I said, I mean we typically don't, well have very formalized way of,
9 ways of working with data. I think the way it typically goes is that we well, depending on
10 the copyright and so on, but I mean like in the simplest case it might end up in a GitHub
11 repository yeah.

12 **Interviewer:** Okay and [hesitates] do you feel like you need better support in this
13 planning phase or are there issues that you encounter?

14 **Participant:** (.) [hesitates] Well I mean, I mean it's not like we have ever felt that we
15 need help, but I suppose that might be partly related to ignorance. I should say that well,
16 I mean typically we don't work with a sensitive personal data or anything of that kind. I
17 mean that's of course a completely different scenario and in that case we would of
18 course make sure that we are dealing with the data in a legal way and so on. I mean
19 usually we tend to work with openly available data sets.

20 **Interviewer:** Yeah and how do you figure that you have, that you deal with the data in a
21 legal manner? Do you ask around or how do you go about this?

22 **Participant:** I mean well, so I mean on the one hand if we use a data set that someone
23 else has created then I mean usually it comes with a license and so on and conditions
24 for use. So I mean that's usually part of getting the data, that we understand how, what
25 are the conditions for using it. On the other hand if we would create the data ourselves
26 then (.) well, so I mean, I mainly work with text data, so I mean if we need to create a
27 data set then I think the best way to go about it is to use texts from sources which don't
28 have, like, copyright restrictions so that we can, well, sidetrack all legal issues and so

29 on. So, I mean, that might mean that we, let's say, work with Wikipedia or something
30 like this, like something that has a permissive license for use.

31 **Interviewer:** I see, I see so during the planning you, when you plan your research what
32 data you use you already look at like "what am I allowed to use?"

33 **Participant:** Yeah.

34 **Interviewer:** Yeah and do you also create a data management plan?

35 **Participant:** No.

36 **Interviewer:** No? Okay. (.) Yeah. Then let's see. [reads from the interview guide]" What
37 do you do with the data in this phase? What tools, software?" (.) Yeah. So, then we go
38 to the next phase. I think the create/collect phase. So, what (--) how do you typically
39 gather the data?

40 **Participant:** (.) So, well, [hesitates] how should I describe that? Okay. So, I mean, like, I
41 would probably say that the most typical way of working for us is that we start with a
42 data set that is public, like it has been created by someone else. Okay. Then, of course,
43 we have to do something with this. I mean, there will be some sort of secondary data
44 that comes out of the processing of the original data set.

45 **Interviewer:** Yeah. And what kind of tooling do you use? Is it just for your benefit,
46 [name]? I think for you, it's probably very natural to say, "oh, yeah, but then I'll just write
47 a simple script to do that". But this is not the case

48 **Participant:** [interrupts] So this would probably be like, well, basically a bunch of
49 numbers stored in big files. So I think typically the secondary data would be created as
50 a side effect of running some scripts. And then there might be other scripts that take
51 these secondary data and well aggregate them, visualize them and so on, whatever we
52 need in order to carry out our research.

53 **Interviewer:** Yeah so and these I think they're processing scripts I guess this is. Which,
54 what languages do you use for example what software do you use?

55 **Participant:** It's mainly Python and the libraries I mean basically it's PyTorch and well
56 things that build on Python.

57 **Interviewer:** Yeah and what's kind of (--) where do you store this data and the
58 intermediary data?

59 **Participant:** So it's mainly, like I think like right now the kind of big data storage points
60 that we have are, well nice storage servers.

61 **Interviewer:** Yeah okay and, okay, do you also turn to support functions at the
62 university or?

63 **Participant:** What are those support functions?

64 **Interviewer:** Well for example, well you're talking to yeah, I work at [support
65 organization] which is a support function we have [other support organization] but
66 there's also the legal department, there's IT services at [university].

67 **Interviewer:** Yeah I mean like, I think that's mainly like, I mean of course we had to
68 apply for access to this storage service like every half year, so we have to submit an
69 application to keep the amounts of terabytes that we have and so on. I mean, day to
70 day, we don't really have to interact with any technical support. I think my PhD students
71 have occasionally been in touch with tech support. Well, for instance, if some, well
72 some version of Python is not installed or some version of some library is not installed,
73 so I mean occasionally, but I would say about fairly low-level matters.

74 **Interviewer:** Yeah. Are there any costs associated with like the collecting and

75 **Notetaker:** [interrupts] Question as a clarification

76 **Interviewer:** Yes go ahead. You're mute, so I can't

77 **Notetaker:** Oh sorry, can you hear me now?

78 **Interviewer:** Yes

79 **Notetaker:** Yeah do you mind if I just jump in and ask for a little clarification with regard
80 to storage. You mentioned that the big projects use like [infrastructure1],

81 [infrastructure2] and stuff like that and do you have (--) how many smaller projects do
82 you have in relation to the big projects and do you store these somewhere else, like
83 maybe do you store them locally or on that sort of thing?

84 **Participant:** I mean like what is a project?

85 **Notetaker:** Yeah that's [hesitates]

86 **Participant:** Every small random thing that I do, I mean like I suppose I have like 500
87 small things on my hard disk, I have never actually tried to count them.

88 **Notetaker:** Yeah okay I understand.

89 **Participant:** So I think like if I just look at (--) okay so I'm the manager of one medium-
90 sized storage project on [infrastructure2] and then the PhD students individually have
91 well their own storage also on [infrastructure2], but I mean I cannot keep track of every
92 little thing that we have on our disks.

93 **Notetaker:** No, definitely. But that helps me. That's great. Thank you.

94 **Interviewer:** Let's see. Did we address costs? Do you have anything that or is this all
95 free at the point of use or is there anything that you need to pay for during the data
96 collection?

97 **Participant:** So (...) I would say in my research field it is not so common to pay for data.
98 I mean there used to be like 10, 20 years ago there were some popular data sets that
99 many people used that you had to pay for, but I don't think that is so common these
100 days so we mostly used, we mostly work with open data sets as I said before. Well
101 maybe in some communities in my field like, I mean if you work with some medical data
102 sets for instance then maybe some people pay for that but I mean in in in our research
103 we generally don't do that.

104 **Interviewer:** Okay, that is very good to know because I definitely have seen some
105 research fields where this experience is completely different.

106 **Participant:** So yes, so okay, so one thing that, so actually one thing now that we
107 speak about payment, so we had one project about a year ago where a PhD student

108 paid for crowdsourcing. So there was a crowdsourcing annotation service that was paid
109 by [funding agency] in order to create a data set.

110 **Interviewer:** Okay, that is very, that is very useful and this crowdsourcing is that like
111 some, is that a commercial platform, is it really open, data to science?

112 **Participant:** Yeah so let me let me try to remember what it was called [hesitates] yeah
113 right, so the service was called [service name].

114 **Interviewer:** Okay, perfect.

115 **Participant:** Let me, (...) right so it's actually [service name] this platform that was used.
116 I don't remember all the exact details because this was a project that our PhD student
117 carried out while on a visit to [other city]. But basically as far as I remember this was a
118 collection of text documents and in combination with each text document, there was an
119 annotation provided by the crowd workers about whether this text was useful or not. So
120 it was basically text together with some sort of human label associated with each text.
121 So then [they] paid the [service name] service in order to get these human annotated
122 labels.

123 **Interviewer:** Yeah okay. And again do you have an order of magnitude estimate for the
124 cost or is that something you don't want to disclose or prefer not to disclose?

125 **Participant:** Yeah okay, so I should probably ask the PhD student but I would imagine
126 that ballpark wise it might have been like 20 000 crowns or something.

127 **Interviewer:** Okay perfect.

128 **Participant:** So [they] basically got like a small grant from [funding agency] in order to
129 pay for that annotation, yeah so I mean these sorts of projects I mean they are not, I
130 mean they [hesitates] happen from time to time in my research field that people use
131 these sort of annotation services or we might pay students or well whatever human
132 annotators are available so I mean we occasionally pay for the creation of data I mean
133 that occasionally happens. I mean, it's similar to, well, the sorts of paid annotation that
134 would be done at [company name] or a company like this.

135 **Interviewer:** Yeah. Okay, perfect. That is very useful to know. Then we move, I think,
136 from the, unless you have more to talk about the data collection process, otherwise I
137 would like to go to the analysis stage, of your data and there's like what do you do with
138 the data during this phase?

139 **Participant:** So I mean like we primarily work on the same basic research in AI and
140 natural language processing so I mean usually the end result of these data analysis
141 would be like we get tables and plots in our research papers, like okay we want to show
142 that well this method works better than that method for instance. So then we need to
143 well run these methods on some data sets and compare how well they work. So and
144 then we present that information in different ways. Like with tables and with plots.

145 **Interviewer:** Yeah and (.) what kind of tooling do you use for this analysis this is again
146 the Python scripts or?

147 **Participant:** Yeah I mean, like for the visualizations I mean we might use Matplotlib or
148 something like this. I mean to compute the numbers I'm not sure well we use any
149 particular tools I mean that would be like, it might be directly based on PyTorch or on
150 like secondary libraries like Hugging Face libraries and so on.

151 **Interviewer:** Yeah and where do you run this analysis?

152 **Participant:** I mean, for large-scale computations, it would typically be, once again
153 [infrastructure1] or something like [infrastructure3] and so on.

154 **Interviewer:** Yeah. Okay. And do you also sometimes run analysis on your own laptop
155 or is the data sets and the analysis

156 **Participant:** [interrupts] I mean, for prototyping, I often use my own laptop or for
157 projects that don't have extremely high computational requirements I might just use my
158 own laptop.

159 **Interviewer:** And are the [infrastructure1] facilities are those sufficient for the research
160 that you're doing?

161 **Participant:** I mean that's basically up to [infrastructure1] policies, I think like I mean
162 sometimes they can be a bit stingy with the number of GPU hours they allocate and the
163 way that they justify these GPU hours and so on. I would say that well, I mean it's
164 basically okay. It's just, well our interactions with them can be a bit difficult, I mean they
165 don't have very clear justifications for their decisions.

166 **Interviewer:** Yeah so if I can summarize that as maybe you would like more
167 transparency on the decision making process in at [infrastructure1] or.

168 **Participant:** (...) Well in summary yes.

169 **Interviewer:** Yes okay, it's of course a very complex and complicated topic but, trying to
170 summarize things here, and yeah so that is so the analysis I guess that's mainly you
171 use the same storage or where as where you put it during the gathering phase?

172 **Participant:** Sorry I didn't get this question?

173 **Interviewer:** You (--) earlier we talked about where you store your data and then you
174 were talking about you typically store this at the [infrastructure2] service or other things
175 from [infrastructure1] and this is also where you use it where you store the data during
176 the analysis phase right?

177 **Participant:** Yeah

178 **Interviewer:** Because otherwise then maybe we can go to the make available phase if,
179 that's okay with you [researcher]?

180 **Participant:** Yes

181 **Interviewer:** So do you make your data available if for example if it's not public data so
182 not of the public sources but do you make your data available?

183 **Participant:** Well, I mean, as I said previously, I mean, for some research projects, we
184 use datasets that already exist, and then maybe that is not so interesting. In some
185 cases where like new data sets are created for a project then we try to make the data
186 sets available. I'm just locating, like yes I just located this data set that we discussed
187 before the one that was created by a PhD student using the [service name]. Yes, so I

188 mean as a general admission, if we create something we want to make it available to
189 the research community.

190 **Interviewer:** How do you make it available?

191 **Participant:** (.) So, I mean, in the simplest case, it might, well, I mean, if it's not a huge
192 data set, I mean, in the simplest case, it might be stored on a GitHub page or something
193 like this. So, in my research community, there is also this Hugging Face repository
194 where many people deposit data sets. So, maybe I can link.

195 **Notetaker:** Sorry, could you repeat the last bit, in your community there is?

196 **Participant:** Let me show you, so here is the link to the dataset that was created by my
197 PhD student, the one that I mentioned before, where we discussed the cost and so on.
198 So this is stored in a repository that is called Hugging Face, where many people who
199 work with AI for language tend to store their datasets there and train models and so on.

200 **Interviewer:** So this is sort of the standard in your field?

201 **Participant:** Yes, I mean, for datasets that can be released to the public. I mean, of
202 course, if there are legal restrictions or financial restrictions then then it might not go
203 here but for open data sets I think it is very common these days that various sorts of
204 artifacts I mean data sets or train models and so on that they are stored in this
205 repository.

206 **Interviewer:** Yeah and Hugging Face is that a like, is that a private, is it an industry
207 party or is this?

208 **Participant:** Well, I believe it is actually a company that runs this service but I mean it's
209 the services are free, so yeah I'm not sure exactly about their business.

210 **Interviewer:** Yeah I'm just the reason I'm asking is for example I (--) maybe
211 [infrastructure4] has some things I know there is for example Zenodo right are you
212 aware of Zenodo?

213 **Participant:** Yeah I mean, like yeah I have seen it occasionally, I mean I don't
214 remember all the technical details about it. That's some sort of system for permanent
215 identifiers or something like this right?

216 **Interviewer:** No that's the ORCID.

217 **Notetaker:** [interrupts] Yeah it does (--) it does provide permanent identifiers but its
218 main function is sharing data openly and preserving it. It's linked to CERN so it has the
219 same funding source so it's sort of like a promise to keep your data for a long time and
220 also give it a DOI number.

221 **Participant:** Okay yeah I mean I have seen these sorts of data sets from time to time
222 that are stored in Zenodo but well I don't remember that we have used it in in our group.

223 **Interviewer:** Oh okay.

224 **Interviewer:** But I mean, I mean there's nothing really preventing us from doing it I think
225 it's just like in my community it's probably just culturally a bit more common that people
226 use the Hugging Face repository these days?

227 **Interviewer:** Yeah. Are there any issues that you encounter when you try and make
228 your data available?

229 **Participant:** I cannot really think of any. Well I mean, well I think that depends on the
230 size of the data, the datasets that we have been publishing I mean they have not been
231 huge. I mean technically there hasn't really been any kind of difficulty to deal with them.

232 **Interviewer:** Just for our understanding, what do you mean by a huge data set, what's

233 **Participant:** [interrupts] For instance I mean, for instance this data set here has five
234 thousand rows, it's a five thousand five hundred rows, as you can see in the deficit
235 viewer so that basically means five thousand five hundred documents, so I mean huge
236 would be like I mean when we start to get into the millions or something like this. I mean
237 then, then well there might be well limitations on storage. I mean I don't remember if
238 there are any limitations set by the Hugging Face repository but I would assume that
239 there are some like if you want to store the whole of Wikipedia or like these large scale

240 web scale data sets and so on I mean then there are usually special technical solutions
241 for distributing these sorts of data sets.

242 **Interviewer:** Yeah, yeah indeed, so I guess the large then we start talking about more
243 than gigabytes with the one that you mentioned here that, this size is about double
244 tenths of megabytes, so.

245 **Participant:** Right.

246 **Interviewer:** Yeah, okay and then related to making available I guess we've maybe
247 touched upon it already a bit is the data preservation because, okay this is making
248 publicly available but is this different from your preservation storage that you can a few
249 years from now still use the data?

250 **Participant:** (...) I mean like, I mean we (--) I shouldn't say that we are extremely
251 systematic about planning for preservation. I mean usually, well our data goes into
252 various repositories and so on and then well we assume that there is not going to be a
253 trade war with the United States that makes GitHub go down for Europeans or
254 something like this, but of course hypothetically that could happen. But well we haven't
255 really given that a thought.

256 **Interviewer:** No. Are you planning on making use of these considerations in the future
257 or?

258 **Participant:** I could say, I could say that this, I mean in the ideal world there would be
259 time for that, but it's not like I should say that, I mean we haven't really given that much
260 thought to be honest.

261 **Interviewer:** No, just to make sure that we are clear on the definition of terms, because
262 what do you mean by a repository that's a GitHub repository or is that something else?

263 **Participant:** Yeah, I think like, I mean like, well okay so I would say that the way that we
264 deal with things practically I mean, like the publication of the data and the long-term
265 storage of the data. I mean, we haven't made a conceptual distinction between those.

266 **Interviewer:** No.

267 **Participant:** So, I mean, if we publish something in the Hugging Face repository or
268 GitHub, that's the same thing as the preservation. But, I mean, like, maybe that's naive
269 of us. I mean, we just haven't really formalized this.

270 **Interviewer:** But this is already very useful for us to know that the availability and the
271 preservation, that those are the same thing for your type of research. It is very useful to
272 know. So then basically the tools that you use are the same as for making available. Do
273 you find any issues, or not really?

274 **Participant:** (...) I mean generally no. Well I suppose, well I mean these cases are quite
275 rare, but I mean like occasionally there can be software issues, like okay, we well
276 software libraries are developed over time and then well occasionally you might
277 encounter some data set that well is not compatible with the latest version of the
278 software so then you have to use an older version of software and this sort of random
279 stuff that happens. But I shouldn't say that this is a very common issue that we have. I
280 mean, usually people use fairly simple data formats like well, TSV files or well, yeah.

281 **Interviewer:** Yeah. Could you explain maybe a data format where you have
282 encountered this issue with backwards compatibility? (...) Do you remember, or?

283 **Participant:** I mean, like, as I said, this is quite rare, but I mean, like, I mean, when
284 people have like, very idiosyncratic format, like, I remember having occasionally
285 downloaded SQL databases, and then spend considerable time just trying to
286 understand like, okay, what are the tables here? And okay, how should I [chuckles],
287 how should I exude this, this data that someone has, well, stored?

288 **Interviewer:** Yeah. But then you're really looking at a much more specialized data
289 format, like indeed, like a SQL database, not like if it's stored as a CSV file or a TSV file
290 or some other more common format?

291 **Participant:** Yeah. Well, yes, yes, yes.

292 **Interviewer:** I'm just trying to get as much information out of you as I can. So I think
293 then we have the final phase, that is the report phase. What do you typically do when

294 you report on the data? What kind of tooling do you use? I think we touched a little bit
295 upon it, like MATLAB and Python. Anything else that you use?

296 **Participant:** So when you say report, do you mean like the publication of the results
297 from the research?

298 **Interviewer:** Yeah, I think that it's (--) Ah, here is [notetaker] coming to a different slide.
299 If you go to the other, yeah, tools report, yeah, that is. It's basically that reporting, like
300 maybe on the data that you've gathered and yeah, the results of the analysis.

301 **Participant:** I mean, like the standard situation would be that, well, we create our plots
302 and we create our tables and then we put them into research papers and hopefully get
303 them published.

304 **Interviewer:** Yeah. And these plots and tables are created with Python and Matplotlib
305 and related Python tools.

306 **Participant:** That would be the typical situation.

307 **Interviewer:** Yeah. Any issues that you encountered there, or?

308 **Participant:** (.) Well not really I mean like except for like, okay how do we make this
309 visually appealing and understandable to the target audience and so on so I mean but
310 well representation problem.

311 **Interviewer:** Those can be issues I've definitely worked with software where it was very
312 difficult to get visually appealing graphs out of it, so it's a valid concern to raise and
313 have. And also the expertise in visual visualization to you have this concept in your
314 head but how do you translate that into a graph that can go into a paper I think might be

315 **Participant:** [interrupts] That's a [hesitates] research problem in itself. I mean how do
316 we present something that is complex so that it (.) so that the take-home message gets
317 through.

318 **Interviewer:** Yeah, okay do you turn to any support for that for example infrastructures
319 like [infrastructure5] for visualization?

320 **Participant:** No I [hesitates] I might have heard that term before but well no I (.) we are
321 quite independent generally.

322 **Interviewer:** Yeah okay that's good to know this, is I think a lot of information for us, this
323 has been great. We have one more thing to ask you, I think first we go [hesitates] no,
324 yes. We have a few smiley faces (--) well one smiley, one neutral face and one frown
325 and the question is for every part of this, the planning phase, the gathering phase, the
326 analyst phase etc. How happy are you. Are you medium, are you happy, or are you sad
327 for the planning phase?

328 **Participant:** Actually I mean to be completely frank, the only sad face that I would put
329 right now is probably in our interactions with [infrastructure1] because of the way that
330 they justify their rejections and so on. I mean, otherwise, it is, I would say, neutral or
331 smiling.

332 **Interviewer:** Okay. I think then we pay, we said, do a sad face at the create and collect,
333 because I think that's where you have the most things. And [notetaker]?

334 **Notetaker:** Yeah.

335 **Interviewer:** Can you put up the frowny face at the create/collect instead of at the
336 analysis? Because I think that's

337 **Notetaker:** [interrupts] Oh, sorry. Yeah, yeah, yeah.

338 **Interviewer:** Yeah. And unfortunately, the format

339 **Participant:** [interrupts] So, I mean, concretely, I mean, the frowning face is more, I
340 think, the analysis part. I mean, like, well, getting access to [infrastructure3] and so on.

341 **Interviewer:** Okay, so it is at analysis.

342 **Notetaker:** I'm just going to create a separate box because it's going to be too iffy.
343 Smile.

344 **Participant:** Yeah, I think my interactions with [infrastructure1], I think there's too much
345 micromanagement. They ask us for very detailed justifications of exactly how many

346 GPU hours are going to be needed when that in practice can be quite hard to predict.
347 And, yeah, and the type of information that they are asking for in the proposals, I think
348 it's not, doesn't make much sense.

349 **Interviewer:** Yeah. Okay. Then, so then we put a frowny face, I think then that is also
350 the, that's the analysis, I guess, for the GPUs. And then [hesitates]

351 **Participant:** Well, I don't think really we have a problem with the services themselves. It
352 is more the interaction with (.). Well, I mean, basically the project allocation.

353 **Interviewer:** Yeah, no, I (.)

354 **Participant:** Apart from that, we don't really have serious complaints. I mean, like, of
355 course, there have been incidents when the service have not been available, but I
356 mean, that's just stuff that happens.

357 **Interviewer:** Yeah, no, but for the format of this questionnaire or of this thing, the
358 analyst wants per block, they want a happy face, a medium face or a sad face and we
359 (.). and what exactly within that block is causing that, that of course is written in the text
360 and is what we now talk about and I think it's been (.). that the sad face is not about
361 everything in the analysis phase but it's mainly about how do you acquire the resources
362 that you need and the interaction with [infrastructure1] in that sense.

363 **Notetaker:** I sort of just threw around a few faces so you want me to change something
364 [researcher] let me know. It's not, it's not very precise.

365 **Participant:** (...) Yeah I think, I think this looks okay.

366 **Notetaker:** I'm guessing that not smiling is sort of more neutral than unhappy here and

367 **Interviewer:** Yeah I have just one question because earlier we established that the
368 making available and the preservation was more or less similar, but I think the make
369 available has a (.)

370 **Notetaker:** Yeah yeah, oh sorry yeah, I see what you mean.

371 **Interviewer:** Yeah and talking of smiles [researcher], what in your dream scenario if we
372 go through this milky way. How, which [hesitates] how would you describe your dream
373 scenario when it comes to data planning, gathering, analysis, and making available and
374 the reporting? Are there things that would make you a lot happier?

375 **Participant:** I mean, I would say that generally speaking, generally speaking, things are
376 okay for us. I mean, I expressed some frustration with [infrastructure1] or, well, the way
377 that they allocate products and so on, which I think I share with my colleagues. But I
378 mean, apart from that, I think it mostly works fine. Like, of course, I could wish that, well,
379 the [infrastructure5] servers were 10 times as big, but [laughs]. But, well, of course, we
380 would want more compute and more storage. But, well, we are generally quite okay, I
381 think.

382 **Interviewer:** Okay. Then I think that concludes the interview. Is there anything you'd
383 like to add before we end?

384 **Participant:** (.) I can't think of anything right away.

385 **Interviewer:** Okay, then I will stop the recording.

386 **Participant:** Okay.