Researchers' Perspective on the Publication of Research Data:

Semi-structured Interviews from Germany

Interview: os_024 – Translation

1	**Interviewer:** Okay. Then my first question is: How long have you been active in science?
2	**Researcher:** For ten years.
3	**I:** Ten years… and what are you currently researching?
4	**R:** Currently I am researching how to develop adaptive learning environments and test and use them. And I am in particular also engaged in the development of co-operative systems. Cooperative systems that make it possible for humans to learn and work together. And in that context learning analytics plays a role in both fields, because there you want, as a participant or observer or researcher, to look at, test and also understand the ongoing processes that go beyond awareness. To this effect also, that you have to do analysis detect for example for adaptive systems: Which groups of learners exist, which representative behaviours those groups show and if the maybe automatically triggered interventions, didactic measures are of use and work and how the resonance and acceptance looks afterwards.
5	**I:** Hm. Very interesting. And with what kind of research data specifically do you work there?
6	**R:** Hmm That is varying. For adaptive systems it is like that, that we for example collect behaviour data from Moodle, gather it and also all of the user input, that comes from the learners there, when they are active in the course. Going further we got documented data, meaning historical data from course occupancy to also identify patterns and groups there, that show certain study behaviours, because we also want to show connections there between study behaviour and the behaviour inside of a course. And we complement this course occupation, compared to a presence university, that there are also events being taken, courses, lectures, internships and seminars. And then we compare that to the learning performance, which was either provided or not provided for the occupied courses, because not in every course an exam is taken. Sometimes the exam is taken years later. Yes, that's the typical data. Beyond that also communication data. Here at the [Institution] a newsgroup system exists and students and teachers communicate in newsgroups and we also look at that data. But those are publicly available though. So they are basically public, because everyone can join this newsgroup. Yes, apart from that, what is missing? Yes, there is also an array of other systems that we look at, but not in the context of the adaptive learning environment but rather different research areas. So in collaborative systems these are usually log data from within the system, that record the user data…outline.

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from Germany. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21644

7	**I:** Aha ((affirmative sound)). And have you already published your research data?
8	**R:** Erm, no. I haven't done that yet.
9	**I:** And why not?
10	**R:** Just because I haven't had the time to edit the data in a way that they are self-explanatory. Although I have recorded and outlined the metadata accordingly.
11	**I:** Okay and…
12	**R:** So that was one reason, but then it wasn't always clear what the right publication body is for that data. Because I… so for the computer scientists as such this didn't exist for a long time. There are big data shops, like Kaggle or something like that. There you can drop them. There they are located between many others. It was always important to me that there is an DOI. Now I know that there are corresponding data journals, but yes, I still have the data lying around. I have it somewhere on the bottom of my to-do list. It is just a question of time. You have to document it properly and maybe you have to do again… so the data is cleaned, but maybe again with different quality criteria. There partly also are some requirements that have to be fulfilled and that one has to check again and clean up for publication.
13	**I:** And would you have, or do you have any concerns in the publication of research data?
14	**E**: No, I do not have them.
15	**I:** You do not have them, ok.
16	**R:** So of course I have to keep in mind that the affected persons can't be identified anymore, so the anonymity is ensured. Pseudonymisation of course. Anonymity has to be checked to which degree some cross connections users… can be limited as a user. That of course has to be prevented. But in my case these are always person related data, thus I have to take these precaution measurements, check it thoroughly.
	[Cut in Audiofile]
17	**I:** … if you work with persona related or sensitive data?
18	**R:** Yes, I work with… consistently with person related data, because they always relate to actions from people, entries from people or can be directly assigned to a person, because these are sociodemographic data. And in the case of exam data I surely assume that the data is also sensitive, because… I don't know… a third party with those, erm, yes, significant information about a human, yes, can find out about others. Sensitive beyond that are the interaction data, but

	like process data from a learning environment, I don't really think so. They would tell quite a bit, but in general only parts of… so data gets collected for relatively short durations of time, so that… oh, are you still there?
19	**[mobile phone sound]**
20	**I:** Sorry. Yes, yes. I am there. That… uhm…
21	**R:** So that only quite small recordings are made.
22	**I:** Aha ((affirmative sound)). And can you publish that data?
23	**R:** I am actually there in the project, that we have here in [City] still in communication with the data protection officer. He did set the condition that… uhm… the data has to be fully anonymised.
24	**I:** Yes.
25	**R:** How that exactly this has to come to pass, he did leave open. But it has to happen now somehow… so when we go that far now, that we also include the exam data, so we have to coarse all further sociodemographic information as good as we can, so that we no longer mention the birth year but only certain age groups or that the residences are no longer relevant, but let's say some identification code for the district. Yes, how exactly we are going to do that we don't know yet. It could also be the case that because of the coarsening of the data we will decide to not include the exam data and course occupancy data in the publication… uhm. We haven't discussed all of it yet and I see it that way: I would publish more data. Anonymised, uhm, pseudonymised they are for sure and the question is now just what kind of coarsening we use and which data fields we completely leave out or which we, yes, change or group in some other way.
26	**I:** And would the anonymised data still be interesting for other researchers?
27	**R:** Yes, I would at least secure that some kind of information can be gained from the data. If that is not the case, you have to ask yourself why you even work on it. So I just say if the coarsening would be that extreme, that you, I don't know, for the study performance only differentiate between if someone did pass or not, then it would be very difficult to relate this to other behaviour data, because I think over 90% or about 95% of the students usually pass their exams. So then you wouldn't have sufficient differentiation anymore. Diligent planning first of all. Therefore that really depends on how you do the coarsening.
28	**I:** Okay. And does the data, that you collect, belong to you?
29	**R:** Yes and no. Me personally, that I cannot quite answer right now, if it belongs to me. I am the

	provider, the administrator of a moodle learning platform of a research entity. Therefore I am the data collecting authority. So any process data, that I collect there belong to me. The user entries, like when they write a text into the forum, then I assume that I'm not the intellectual owner of their mental output, but that the student is that himself, him as the author…I can maybe collect some key figures on those entries, that then again belong to me, because I have deduced them. Concerning the exam data, they belongs to the examination office, as the examination office collects the data. Or, what would be an exceptional situation, that my boss, who also holds the exams in this subject, because he collects the original data, writes down the grades on a form. There it would at least stay in the scientific field, but that would only affect a small part of the exam data, of the documented data. They belong to the study administration here of the [Institute], so that is just based on good will, so when they are well-meaning towards us. When you can sell it them well, that you want to use the data for a meaningful cause, then you get the data. But the ownership is really diverse in this case. So it depends on the individual data.
30	**I:** Would you even have the right to assign rights of use and to publish the data, when it's so complex regarding the ownerships?
31	**R:** There is still discussion to be had, but there is a research focus, that strengthens my back, that covers such cross-institutional questions, because the data were collected within the [Institute] and if I then would not have still the opportunity to publish them. I think there are surely some decisions from the side of the rectorate needed, to in a way establish clarity. But of course I have to make sure, before I can make a publication like that. When I include the data, if I am allowed to do that.
32	**I:** And do you have the feeling, that the process of the publication of research data is complicated or too opaque?
33	**R:** It surely is complicated. Uhm, complicated firstly just because you have to obtain certain data first, because it… it has good reasons why they aren't just lying around openly, that you can just use them, because they still have personal reference and we connect the data that the university collected with the enrolment of the people for example and the data that we collect in the learning environment. Those connections we have to make. Thus, we can't use the anonymised or pseudonymised data of the administration, because then we couldn't connect them anymore. And when you then want to get this non… not yet pseudonymised data or when you want to task an independent authority with this, that this data gets pseudonymised in a certain way and with a certain algorithm, then it gets very complicated. You have to talk to a lot of people and the same way, especially the coordination with the data protection officers adds to that, when then again it's about publication. I think the whole thing is the easiest, if you conduct your own study and survey all data yourself. And in the consent form, that the participants of the study sign, when you just note that a publication is planned. The purpose, that is immediately revealed for the data usage and which they agreed to, then it is quite simple, but that would in my case always just be a

	small study, where I could easily do that. But with all things affecting a larger number of students and I talk about student numbers in four digits, it is a bit more complicated.
34	**I:** And what do you think which information would be needed or which information would have to be provided to researchers, so that they publish more data for the purposes of Open Data? What would help you to publish more?
35	**R:** Well, I could clearly imagine that there should just be more like some kind of manuals, in which it is explained step by step what I have to do and which tools could be helpful for me in this regard. So what I mean, uhm, that begins again with the data cleaning. So you can exercise even more care for the publication of data, to search for some outliers over again in the data. And if that affects areas that you yourself haven't even evaluated or something like that. Uhm, it's even about small blank spaces, that the statistics software benevolently ignored and correcting them, but that… in cases like that the data cleansing is a bit… as far as to questions: How do I even describe a data field well and who tells me, that it is now understandable. And with the review, well, I have no experience with that, if… how well a review from a specialist colleague is carried out and how much it is worth after all, how well can someone then put oneself into this. Yeah, that is still just difficult. And of course the question remains: Huh, where do you publish that? And I have already received here some tips and I already know some places. But there is still a insecurity there. Where am I best placed with my specific survey results? In the sense of: Where do I get found best? Where do similar data sets exist, that… which I can maybe also use as orientation in the way how maybe I myself have to document my data. And there is a bit, if you research interdisciplinary, you don't know exactly where you're looking. So I am now with the humanities and social sciences and publish my stuff there or do I go to the psychologists, where I also don't really belong there as a computer scientist. That isn't quite clear. Easier it would of course be, if every journal, every conference just had a data… Data Shop and data store or something like that, where it would be natural: Yes, there the stuff has to go! And if multiple conferences, let's say of the computer sciences or something like that would get together and create such an institution, then it would in my opinion still be better than what the universities offer as solutions. Or the things are also not being found.
36	**I:** Hm. Very interesting. I find that very helpful. Thank you for that so far. Do you know if research data in your scientific field is published more or less in other countries?
37	**R:** Well, there is anyway a distortion in (unintelligible) bias. There are significantly more publications from the Anglo-American area than the German area. Just because the language community is bigger and with the data publications it should be similar.
38	**I:** Okay. Great. That was actually my last question. Thank you very, very much. Thank you.
39	**R:** You're welcome.