Researchers' Perspective on the Publication of Research Data:

Semi-structured Interviews from Germany

Interview: os_011 – Translation

1	**Interviewer:** First of all thank you that we can hold this interview.
2	**Researcher:** Yes, gladly.
3	**I:** And firstly I would ask you to tell me in which areas you do research and what you're doing.
4	**R:** Okay, so I'm active in the areas of learning analytics and educational data mining. We have two core topics that we deal with. One being basically how we can use the procedures of machine learning to, in the best case, generate online courses automatically. And on the other hand we deal with personalisation, so direct usage of learning analytics, where we try to use different information, to survey for example personality features or cultural information and use them like that, so an online course can adapt to a person, so that they can learn in the best way possible.
5	**I:** Very interesting. And with what kind of research data do you work there specifically?
6	**R:** We have our own data, as well as data that we get from partners or directly, when we for example go into the area of online course generation. Maybe one of the branches is question and answer generation based on texts. There we have different databases that already exist, where the texts are already annotated and such data is already available, but which are not personal, but only content-related.
7	**I:** Aha ((affirmative sound)), okay. Do you also generally have personal data that you work with?
8	**R:** We also have that. Uhm, that mostly when we look into the direction of personalisation. There we survey personal data. With questionnaires for example. With cultural questionnaires, personality questionnaires, in order to then, yeah, survey some data about the person in general. On the other side we also collect like clickstream data, so user behaviour within online courses, to bring that together, with the cultural dimensions or the personality traits, to just find the connections there.
9	**I:** And have you already published research data?
10	**R:** Erm, only aggregated. So really just as an evaluation, because fully making the data available, as far as we know, is not possible, because there is a lot of personal related data present. Especially when we survey cultural dimensions it is like that, that with the few data points that we have about the person you can identify the person clearly.

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

11	**I:** Even after the process of anonymisation?
12	**R:** That's rather problematic, when we look at when we work with cultural dimensions, then we got for example five dimensions, that we view and in each case we have the characteristic. But then if we connect them as a basis, then it is hard to anonymise them, because we have different parameters, that we have about single persons. But we need them, to then carry out the next steps, like the personalisation. To train machine learning you need exactly this data and can't just change them in any way.
13	**I:** Is it, in your opinion, then even possible to publish data in the field of learning analytics? Would there be any other kind of research data, that would make it possible?
14	**R:** So in the first place is learning analytics by definition a very large field. And when you now look: Generally, in the field of learning analytics I would say, yes, you can publish. When I now, for example, look at the clickstream behaviour of people, then it is per se not necessarily bound to the individual person. There I would state, yes. There exists for example this data shop. There is already a lot getting published. As soon as you move into the area of user modelling and then build user specific models, it gets hard to publish those, because they are indeed only related to one person. And I know there is a new thing in some kind. So you can for example, if you, erm, build neural networks, then there is this fake neural network, hm. There you should be able to transform the data, which you use as raw data, in such a way to make it publishable, so they are supposed to be get anonymised. I don't know if that works. Only heard about it last week. Though I can imagine that it works.
15	**I:** I need to research that myself. I do find that very interesting. And have you perhaps heard of the FAIR principles?
16	**R:** Uhm, no.
17	**I:** So findable, accessible, interoperable and reusable.
18	**R:** Okay, sounds logical.
19	**I:** Indeed. It is about the publication of research data and according to these principles it is also said that you don't have to publish the data per se to create the data fairly, but maybe just the sets of metadata, so that other scientists know what you yourself… what you are doing research on or where co-operations could be developed.
20	**R:** Okay. So what I always especially do when I train neural networks, then I do or I always write in the publication how I optimized the hyperparameters, what I basically… it is a very long process until you find the optimal parameters and they will then be actually published. So not the pure raw data, but rather… I basically take data, train the neural network and the many tries that I did and which… and this computer has to actually calculate for weeks to find the optimum. And these

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

	parameters, that were identified there, I publish then, so that someone who also wants to run that experiment again doesn't have to go through this step once more, to optimize everything per hand, but can just do the next step which is looking at the results of their data.
21	**I:** Aha ((affirmative sound)). And, hmm. You said that you think, that in principle you can't just publish the data. There was a lot of uncertainty within that.
22	**R:** Of course.
23	**I:** What could be done to take that uncertainty away from you? Meaning what kind of information would you need to simplify the process of publication of research data?
24	**R:** Erm, good question. So in the area of data protection I'm not quite inexperienced. I have been already able to gather some experience with that. I think, I would rather answer this generally, as well for the others in the field of learning analytics, where I'm also in exchange with other scientists. I think the problem is that many can't differentiate between what are these sensitive person-related data and what are the data, that are initially not related to the person. The second problem is actually that the connection is often there or has to be there between personal data and the non-personal data, because that often is the core competence of learning analytics. To only publish one of those often doesn't help anyone. Unless for example I go into that realm where I say that I want to generate online courses where I don't have any personal data at all, there it is relatively easy. And there everyone knows: I can publish this. That's fine. But as soon as I have personal data in there, it is difficult. It would maybe be easier… or will get easier in the future, if there were real guidelines available, infographics, where you could see at one glance which data is sensible in what degree and if those could be published within the data protection regulations or not.
25	**I:** And you said that the data for the online courses could be published. Are they actually published?
26	**R:** Depends on which data. So the question is… click stream data?
27	**I:** Exactly. The non-personal data. Do they actually get…
28	**R:** They also get, indeed, published. Yes.
29	**I:** And do you have a sense if in other countries in the field of learning analytics there are more or fewer publications than in Germany?
30	**R:** Significantly more.
31	**I:** Do you maybe have an idea where?
32	**R:** Especially in the USA. The example of Data Shop, which is basically the most known in learning

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

	analytics. There are lots and lots of data getting published there from different, yes… from different studies and procedures and there is indeed personal data getting published, because probably there is no deep importance attached to personal… or the sensitive personal data. Yes. That as a thesis.
33	**I:** Yes. I like that. Would you even be allowed to publish your own research data? Not in the sense of data protection, but do you own the data?
34	**R:** Uhm… That is a great question. So I would say the data does not belong to me. Everyone, who basically… no, everyone who participated in a study produces data and this data then also belongs to that person. Uhm, yes. That easy. That way I would phrase it and I think it's also written like that in the data regulation law, the basic data regulation act.
35	**I:** Yes and do you know if your institution has a research data policy?
36	**R:** There is one, yes.
37	**I:** Very good. Okay. That was actually my last question.
38	**R:** Oh.
39	**I:** Thank you very, very much.
40	**R:** With pleasure.

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