Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China

Interview: os_006

1	**Interviewer:** Okay, seems to work. So again, thank you very much that you are willing to have this interview with me. It's very helpful for my research. So, in the beginning I would like to ask you to introduce yourself, where are you working, what are you working on, what is your research about?
2	**Researcher:** Okay. Yeah, sure. So, actually I finished my undergraduate studies in [China] and then I went to United States to start my PhD study and I got my PhD degree in computer science from [the university]. My PhD advisor is [name and description]. So, in my whole PhD study I am kind of working about the intelligent tutoring system and also looking at the (unintelligible) and artificial intelligence in education, this kind of area. And I forgot, I came back to China and joined the [university] and spent two years there as a postdoc researcher and also a research scientist. And then I came to [the university] and currently I am a associated professor, in the same centre as professor [name], yeah. So, yeah, so basically, I'm kind of a mixed, how do I, mixed experienced in United States and China.
3	**I:** Mhm ((affirmative sound)). And what are you researching on? What is your research about?
4	**R:** My research… generally my research is about how artificial intelligence can help students learn more effectively. And also, how we can use artificial intelligence or other computing technology to mime students' behaviours and discover some efficiency through the learning behaviours and inefficient student learning behaviours; and how we can use technology to design some automated agent who helps students to learn some deep knowledge, something like that, yeah.
5	**I:** Very interesting. And how long have you been working in science then in total?
6	**R:** Ah, you mean, you mean in this kind of area in total? Or//
7	**I:** In general. I mean probably from the time when you started your PhD I would say it's the time where you//
8	**R:** Uh, it's about nine years. Yeah.
9	**I:** Mhm ((affirmative sound)) okay. And what research data exactly do you work with? What kind of data?
10	**R:** Uh. I would work with program research data. So, for example for my PhD study, we actually do some experiments based on subjects. So actually, we call some students, kind of we call some

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21330

	students and pay the subject fee to the students and they volunteer to, erm, to participate in our study and uh… and yeah, so. I don't know whether you know that, in the United States there is RD process and they actually go through the RD process and sign some consent form and we collect the data and we work with that data. And also, by the way, we actually developed the system by ourselves. It is called//it is some intelligent tutoring system by ourself. And we use that system together with the student behaviour data and log this//erm, the format is log data. And in… we also actually get, uh… gather some sensor data from the students, like the chair sensor, skin conductions and also the camera//camera data and the mouse//mouse pressure sensor. And so, we have the log data and some sensor data and we worked//we started to design and implement some agent, computer agent, to help the student learn. And this is within the United States. And in China we, in specific in [university], we kind of more//more likely to work with the data from, kind of from the government. Because we, actually our centre in the [university] will implement a system called [name], but basically, it's a learning management system, so the students actually do the//do their homework and also with some kind of MOOC//(unintelligible) in the platform. And also, the//when the students do the exam in the physical classroom and the teacher sends the exam data into learning management system, so the system basically has some student homework data and also exam data. And, erm, also, the student can do some one-on-one tutoring in the//in the platform. So, the//it's kind of a… yeah, probably I'll need to explain in some more detail, because it's so complicated here. So, you know one-on-one tutoring, the student initiates the one-on-one tutoring station and there will be actually in the course some school teacher behind the mirror and the teacher can have students one-on-one online. And they also gather that data, erm… about their tutoring stations and they can also be working with this data. And they combine the one-on-one tutoring stations data and also the students… the data in the learning management system together, so we know how the students' academic performance changed. So, but the data owned by the... actually owned by the government because all the students went to the public school owned by the government, so basically the data actually is owned by the government and because we are the [university] working with the [city] government so we can… they have the access to the data and... Erm. Also, we do the analysis on the data also try to have the improvement on the platform. Also give some feedback to the government and let the government know how some voluntary surveys work and how the school's performers generally work with the learning management system. So, we also use that data. And the last side of data we are working with is that they gather some students' short answers and we build a (unintelligible) model based on the students' short answer, but to do that they also need to have a lot of so-called labelled data. Is//is erm the labelled data means that they need to hire some school teacher to manually grade the students answer, but they actually… to have the uhm acceptable agreement they need to hire two teachers to grade the same students answer and calculate the (unintelligible), something like that. And uh… so, yeah, so basically in that case they gather the data from the school teachers and also get the data from the students. So yeah, generally basically the data we are kind of

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21330

	working with currently and before.
11	**I:** It's a lot and very interesting. You said that the data kind of belongs to the government, right?
12	**R:** Mhm ((affirmative sound)). Yeah.
13	**I:** So, are you allowed to publish the data you are working with?
14	**R:** Uh. Currently we haven't figured out the way to publish data, because one crucial problem is that when the student use a learning management system they need to input their real name in the system, because the teacher also need to check the students' record every day. So, this is a requirement to// for the student to input their data into the system. And when they do the analysis, they anonymise the students name, but… uh… but they cannot find a proper regulation, a rule for publishing the data, because also, well, they can't publish the data, they need to confirm with the government or confirm with the… I think also… we are not sure whether like we have the... very appropriate consent form to the students or their parents, so this is the issue. So, yeah, yeah, so we haven't figured out where to publish the data.
15	**I:** Okay. And so what information would you need to make this process of publication easier or even possible? Is there something that could help you to publish the data? What would it be?
16	**R:** Yeah, I think we really want to publish the data and we… Actually, I think the first thing we need is a very clear guideline about how the data can be published. For example, I also attended some workshop about the ethics of the data. So, some people say, even says, anonymisation is not enough, because you can figure out what the student is by looking at the column data and then combine them together. So, for example if you have the age of a ano//anonymised student and the score of the student and also the sec//the section, I mean the, kind of the physical section of the students. And sometimes you can figure out a way to... to find out who exactly the student is. So, so like... I think we need very… we really need a clear guideline about how to do the anonymisation and to protect each student. So first we need a guideline… a step by step guideline how to publish the data. And also… I think there is another one problem, is there any like, tool or platform to support this kind of process. Uhm, like... Because if//if the step by step guideline is just a word document or pdf it's really hard to follow//follow that and... So, if there is a platform like, uhm, something like if you submit a paper there is a platform how for you to, uhm... specify each item I need to submit the paper. So, there is, I think there is a platform required for that one. And, uhm, I think the last thing is... that if we want to publish the data, we really want other people can like… make use of the data. Otherwise there is a million ways to publish the data. So probably it is... it is very demanding to have a kind of uniformed data format, so… or some syn//synthetic requirement or specification of the published data. And so that other people can like reuse the data easily and can understand the data easily. Because for ourselves even when they are work//working with the learning management system data they cannot have a lot of incomplete data and some column is missing and some part is missing. And

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21330

	for some part, we… even for us, we don't know the, like, the real meaning, because the ones who implemented that have already left, so there is some really, really trivial things here, but it is really craft costing here. So, I think it is really important to have kind of a uniform, a data format, so that people can have an agreement on the different data, so that really it is meaningful to publish the data. Uh, I think, like, the main... gather the student data, for example we look at the things like xAPI and they actually describe how the student behaviour, learning behaviour should be recorded. And I think something like that is, like, need to embedded into the platform to have, to publish the data some… and to have the reader really understand the data. Yeah.
17	**I:** Mhm ((affirmative sound)). Then I have a question: How it works in China? If you publish a paper that is based on data obviously, on research data, do you have the possibility to provide the data as well or do you only publish the paper without the research data it's based on?
18	**R:** I think it's just the option for us to publish the research data or not. And, uhm, yeah, we always can like provide the research data as an attachment or some weblink here to, uhm... to kind of support the research paper. And, yeah, I think it's just like other, kind of like other international papers, yeah.
19	**I:** Okay. And.. do you think in your discipline is research data more or less published in other countries than in China?
20	**R:** Uhm, do you mean other disciplines? Compared to other disciplines, or//?
21	**I:** To other countries. Like in Learning Analytics or for example, I don't know: If Germany publishes more? It's only on your feeling. What do you think how it is?
22	**R:** Uhm. I... don't have a strong feeling about that, but I think it is maybe… in the United States, there is a, like, platform for the published research data like data shop, owned by the CMU, Carnegie Mellon University, they have a data shop and there is a lot of learning analytical research data there. But I don't think China have anything like that, but yeah. The one thing I remember is that, uhm… I don't know whether you know XuetangX? XuetangX?
23	**I:** No, I don't know.
24	**R:** Basically, it's a MOOC platform owned by Tsinghua University. And I know they are about to publish a lot of MOOC data recently, but they haven't done yet, but they are kind of in their schedule they have already said that in public. And so, yeah, I think this is one thing I know, I'm aware of and, uhm… Yeah and other than that it's a… Yeah, I think that... I think it's always good to have more people to publish the students learning behaviour data.
25	**I:** My last question is: Do you ever… Have you ever thought about licensing of data? Like when you think about publishing or, yeah, sharing your data? Have you thought about the licenses that you could give to other researchers? Do you know some that you could use for data?

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21330

26	**R:** Hmm, no. But I think for the, like for the practical purpose, so usually they want to use the data. When they gather the research data, they want to use the data to publish some paper and probably after two or three years, or maybe at least one year maybe we they will consider to publish the data. But for the licensing issue, I haven't thought about that. Yeah.
27	**I:** Mhm ((affirmative sound)). Okay, that was already my last question, because you answered the other ones before, which is really great. Thanks a lot.
28	**R:** Okay, yeah.
29	**I:** That was a huge help for me. Thanks a lot. We can keep in contact. So, I can tell you about my research results later. So, thanks a lot. If you have any question afterwards, just let me know. If you know other researchers who want to be interviewed, I'm very happy to interview even more, so…
30	**R:** Okay, okay. Thank you. So, I'm looking forward to seeing your report about that.
31	**I:** Oh yeah. I'm very curious too ((laugh)). Thanks a lot. Have a nice day.
32	**R:** Yeah, bye.

Biernacka, K. (2020). Researchers' Perspective on the Publication of Research Data: Semi-structured Interviews from China. Humboldt-Universität zu Berlin. DOI: https://doi.org/10.18452/21330