Cite as: Kalová, Tereza. (2020). Metadata for Research Data: A Needs Assessment in The Sciences Interview Transcript Dataset [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3897321

I: Interviewer

B: Befragter / Interviewee

(unv.) = unverständlich / unintelligible

1. **Wissenschaftler E – Teiltranskript, persönliches Interview (30.04.2019)**

2. […]

3. **I:** Could you please just introduce yourself and describe your area of research, briefly.

4. **B:** Mm hmm, I'm [a senior researcher at a University in Vienna] and I work with the physics most broadly but more, more specifically materials science and even more specifically nanomaterials such as carbon nanotubes and graphene. And … our main research technique is electron microscopy to generate research data. But also um first principles simulations […], where we do generate also quite a bit of simulation data.

5. **I:** Mm hmm. Alright, so how do you define research data in your discipline?

6. **B:** Um, well I guess the most broad definition is any data files that are the outcome of a measurement or a simulation. Um, but. So, we deal with research in kind of two levels. One is just at the raw data that we generate every day. That, let's say 50 percent of that may not be useful at all. We can immediately see that this is not useful data. Um same with simulations. There we run simulations, sometimes they fail, sometimes the computer crashes, whatever. There it still might be big data files

generated all the time. All of that in principle is research data. But in terms of our data management. So for the ERC grant, I had to submit a data management plan for how we deal with data and in that more narrow definition, we talk about research data as data that is directly tied to publications. So the open data in that sense is for us publication-centered. We want to have the paper, the article that describes the results but also to have that data available as open data. And that's kind of the more "eingeschränkt" definition that we use mainly for research data.

7. **I:** Right. So does the term metadata tell you something.

8. **B:** Yeah, sure. So, metadata are the descriptors you give to the file about what is the provenance of the file. Um for example in our case, we denote for electron microscopy images, we denote the instrument, where it was measured and some small details about the experiment. And for simulations, again the code that was used to generate the data, stick to stuff like that.

9. **I:** Mm hmm. So could you please describe what this metadata is for?

10. **B:** Do you want the naive answer or the informed answer?

11. **I:** Both, perhaps? Or whichever one you want to give?

12. **B:** Yes, I know a fair bit about text and data mining and machine readability and interoperable metadata standards, and repository interoperability and I've been following this Open Data space a bit. As a practicing scientist, metadata for us is that you need to, when you have a data file you need to somehow understand months and years later, what this data file is about and the metadata is the context that allows you to understand the file.

13. **I:** Alright. Could you please describe the role metadata plays, when you're working in a team, when you're collaborating with others?

14. **B:** Um … Not so much, I would say, not really. Metadata for us is more about archiving. When you are actively collaborating, you're looking at the files together, you are discussing them. Um you're creating a figure for an article, metadata that doesn't really come into it much. It's more about archival and Open Data purposes that we use metadata. That's, there is kind of the formal metadata, of course, when

we take an electron microscopy image. So that's our bread and butter research. There is a lot of metadata that's gets automatically saved about the state of the instrument. And sometimes we go back to the metadata and check some parameter later on. But typically that is getting automatically written, getting automatically archived but we never really look at it. Same goes for simulation data. There is, for most of the data files you can see, what were the parameters of the simulation that generated this for this file. But it's not something we necessarily refer to during the research process itself.

15. **I:** Do you ever reuse data from others? That were collected from others at that other institutions perhaps or just from colleagues.

16. **B:** Um. Certainly within the team. Yes, it's a lot of sharing for collaboration, it's typically not the raw data that we share. We are working on a specific publication and we want the specific measurement or some simulation that answers a specific part in the story of the paper and then that usually is more of the refined and processed and analyzed data, not the raw data itself. But in terms of Open Data, I don't think I've ever used other people's data. I've offered our data to be used. I'm not sure if everybod/ anybody has. But yeah, it is out there.

17. **I:** But you have never really done that yourself?

18. **B:** Just taking a big raw data set from somebody. No, not really.

19. **I:** Or not necessarily raw data. But maybe even just …

20. **B:** Well an image file yes certainly, yeah. If for example, quite often I'm writing the paper and then I collect figures from other authors. Sometimes if the entire figure is coming from one collaborator they send the whole thing like typeset and layout. Other times, they send the, if not the very rawest form of the data still the relatively unprocessed form of the image. And then I read that in, perhaps to a specialized software, I may play with the contrast, copy something and then put it into our layout software and make the final figure. So yes, I have in that sense dealt with data files from others.

21. **I:** […] So has metadata played any part, when you're perhaps using an image from someone else? Or is it just that you get it, perhaps from a colleague and it's very

informal. How, could you please describe that?

22. **B:** Don't think I've ever actually looked at the metadata of a file from another collaborator. So, I trust that they send me the correct measurement and take, then tell me in an email this was measured in these and these conditions. Sometimes it's not necessarily written in the file and of course, yeah. So I don't think I have looked at metadata to get this information.

23. **I:** So in which language do you describe your research data?

24. **B:** Um, like English or do you mean a metadata language or specification? I forget, what the standard is that they use on the [institutional] repository so I have open data sets on [the institutional repository at the university] and I've used their classification schemes and metadata schemes. I forget now what the name of it was. Normally for us metadata yeah is just English, common spoken language description. This was measured here and here and the microscope was operated at this voltage and perhaps the electron beam current was this and this much, it's very descriptive in normal language.

25. **I:** What are the reasons for using English?

26. **B:** Uh just cooperation so. I mean German is not my native language and we have some people who don't speak German at all in the group. Um very often have students or postdocs, who come from abroad and in the beginning they don't speak any German. Plus we hardly have any German speaking collaborators. It's very international, so English is the common language.

27. **I:** Um, so have you encountered any problems when writing today when describing your data in general?

28. **B:** Yes, it's laborious. So partly it's a user interface issue but [in the institutional repository] for example you have page after page you have to click through and in most pages, they are not at all relevant for the type of data. There is all the artistic metadata pages and everything we just have to click through them, they're not relevant for us, but they're still there. And even just filling in the minimal metadata that's relevant that still takes for each data file or data item that takes relatively, I mean it's not a long process maybe five minutes or three minutes or five or

whatever, but for doing a lot more Open Data, it's still a bit too much. It would be nice, if it was more streamlined. So, it's not difficult as such, you might have to check the meaning of certain fields in the metadata. OK, what do they mean by that? But it's, you know, it's not too difficult. It's just the manual and laborious way of it. That's slightly a hindrance for doing it more.

29. **I:** Right, so what would you prefer? What do you think would make this work easier?

30. **B:** Yeah, good question. Um it would somehow have to be a direct integration between our measurement software and the data storage. So we are doing in addition to this Open Data on [the institutional repository] we have our own data management system. It's, I forget the specifications, but it's a RAID array of disks that are mirroring the contents, so there's some data security and from our measurement instrument everything gets put on this file server once per day. And of course all the automatically written metadata gets also automatically written there. But the software, while it is open source, it's not in no way a standard software. So, the files that it writes are not usable, something that you would put as an Open Data. So when we want to share it with other researchers we have to convert it to another format and then write the metadata by hand basically currently.

31. **B:** In principle so the way we. Our files are basically images, black and white images of electron microscopy measurement. Uh in principle, I guess it would be possible to write the metadata from the software into that. So we use TIFF tagged image format as our most interoperable Open Data format, but the main metadata the capabilities of TIFF are very limited, so we could somehow hack it, that it writes it into the metadata but it wouldn't be really readable for most other people. So um it's a bit of a problem with the data format. There are some other formats that are in use but they are proprietary and we don't actually use the software that mainly other labs use. So yeah. It's, I don't really see like how the workflow would look for (unv.) for all of this. It's currently not really clear.

32. **I:** So you mentioned that you use [the institutional repository]. Can you please imagine now that perhaps you complete your current research project and you want to share it with others. Perhaps just colleagues, perhaps everybody. How would this be possible, if somebody else wanted to use your data?

33. **B:** Uh, well it's not something one should consider at the end of a project. Yah, it sh/ I think. I don't know how the new Open Data practices will develop in the future, in Horizon Europe for example um. Currently I think the publication centered one [Unterbrechung]. It's more about, to me about reproducibility and about giving the, ‘cause very often when we show an image in a paper, we have to make sure that it's very clear, when you print it and so on. So, we have to process the data in some way and it's somewhat subjective process. So, having the raw data file available as Open Data, I think it improves the research product, the article itself. And that's the main form we concentrate on. And this we do on a rolling basis whenever there is a publication.

34. **B:** Um for long term archival uh type, of course that was part of the data management plan. So, for that we will have just our own service, where we store this. It's frankly so much data that … And so small part of it in a routine day when you are doing electron microscopy, you might record 200 or 500 images and maybe 20, 30, 50 whatever depending on how good your day is, are actually really useful and valuable. And it's not really possible to, yeah. Look at just the entire dataset and figure that out. You actually have to have the context of what you were trying to measure and what are you looking at at each moment. And that's not always clear from just the data. We do have an electronic lab notebook, where we do try to describe the running of the experiment and what we were trying to do and that can be in our system, can be correlated with the data files. But it's not something that we, we haven't considered doing the like open notebook science and sharing this real time. At least not yet.

35. **I:** I see. Are you aware of research data management training or consulting services?

36. **B:** I did one course. Research data management at [a University] or something like that was the title. It was a half-day training. And that's where I met some of the people from [the institutional repository] and actually they, yeah. They implemented some features. That … Well not the whole. We had some feature requests, they helped with some of them but some of them were too technically demanding for their specific use case. Um yeah. And … I know there's consulting companies and other things like that, who in principle help but that data management plan that was required for the ERC, it was an afternoon for me to write. So, it's not that demanding. If you have big consortium projects with multiple partners and industry

and confidential data maybe patient data, human subject data whatever that's a lot more complicated. And there, yeah having professionals might be required. But for this very kind of simple research this wasn't that big of a deal.

37. **I:** Right. So, what was your experience with like this course that you took at the University […]?

38. **B:** Yeah, I found it. It was good. I think it was nicely compact. I've done some other trainings, where it's like two days and I feel that the content would have been just half a day. And for me, where I go somewhere for two days, it's really a big investment in time and I would really appreciate the condensed form. I think this was pretty concise and yeah, very knowledgeable people and um good to actually meet them also because yeah. Sometimes we do need to ask for something that we want to do with [the institutional repository]. I was using Figshare before for an earlier open dataset but there's not really a big advantage, you can get the DOI from both services. And yeah, we've tried to move from external services to [university] services whenever possible. Like I was using Dropbox before for cloud storage, now I'm using [the university software] and so on. And [the institutional repository] is, it's a fine repository, it fulfills our purpose quite well.

39. **I:** OK, that's great. So actually, we've come to my last question. […] Imagine everything is possible. Finances don't play any role. And you can imagine whatever would make your work easier, when it comes to research data management and especially the metadata. What services or support could the university offer you?

40. **B:** Um … Yeah, I mean I did send my, so, many projects like this, I think they're fearful that also require this if not yet then in the future, just getting some help in writing the research data plan. And I actually did receive this after meeting with the people [from the institutional repository] at the training, I sent my draft of my plan to them, they offered some comments and feedback and then I modified it slightly. So that was already helpful. Yeah, I think it's probably true for most physics research, most natural science research is that the data is generated in so many different ways that are very specific and no generalist data management support person can be familiar with all the technical nuances. So, I think it kind of has to be in-house. What we have is a very lucky thing, is that we have one staff scientist, who is taking care of the instrument, the microscope but also taking care of our compute

clusters or our storage clusters. And he's very, he's a doctor. He is he was working here at the […] as a researcher before and then he became a staff scientist.

41. **B:** And if he wasn't working kind of half-time on the data management internally, we would, yeah, it would be very difficult. So in terms of the university, yeah, staff scientists, that would help with everything, from instruments to data. But uh yeah. I appreciate that the help desks and the trainings and [the institutional repository] and so on, but I think it's very often so specific, it needs to be within the team or at least on the level of faculty or we have these research groups, the Physics of Nanostructured Materials, which is actually for professors or something. At least on this organizational level, there needs to be somebody, who is knowledgeable, I think. And very often these are very hard positions to find.

42. **I:** Is there anything else that you would like to add to this subject?

43. **B:** Yeah, I expect that I have more knowledge than 90 percent of people on this topic. So please (lacht), don't take my level of knowledge ability to be the benchmark. ‘cause I had to very recently face this and really think about this a lot. Plus, I've been very active in the Open Access and a little bit Open Data side and I was evaluating the Plan S requirements for publications and repositories and so on. So, I went through all the technical protocols that they're asking for. So, I've been just doing a lot of my homework recently, so.

44. **B:** But … I think it's broadly awareness is coming and now that more funders are requiring, I mean ERC would have allowed an opt-out but since, I mean that might have been the easy way out just less work for me, but you actually do have to opt-out if you want to so. I understand Horizon Europe it might be a mandatory in anyway. So, I think there is more awareness coming off these issues but, yeah, I think still most people are not fully aware.

45. **I:** Can you imagine that data management being part of the curriculum for students in the future.

46. **B:** Yeah. So we do have scientific computing and programing and so on and I’m, I don't teach it myself, so I don't know exactly what's in the curriculum but I think data is discussed at least to some degree. Mm hmm. Yeah, for young PIs, if you get your own project then this should be training basically mandatory, or at least a

strong recommendation and. But for me it was available, it was nobody really pushing to it but I knew that I had to write the plan. OK, check the course program, there is a course, I take it. So I think it's currently available but whether it's known enough and attended enough, this I don't know.

[…]