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 C – Teiltranskript, persönliches Interview (29.04.2019)**

2. **I:** […] Can you please introduce yourself and your area of research?

3. **B:** Alright, um. My name is […] and I am Diplom-Biochemiker. So and I, um, my area of research is cell biology and biochemistry with a focus on advanced microscopy methods, um, physics you could say and studying signal transduction process. […]

4. **I:** So, how would you define research data in your discipline?

5. **B:** Research data, uh, is basically the results of empirical observations, which are either collected using an operator or using machine. And this is, this includes but probably not limited to electronic data. Our lab books entries are the primary data, include primary data as well as the secondary processed data. So, I don't know if this suffices? Good, yeah.

6. **I:** Does the term metadata tell you something, does it ring a bell?

7. **B:** Yes, it does.

8. **I:** So how would you define it?

9. **B:** Metadata are basically the way to describe the common properties of the data, which probably would include several specific things like the operator and as well as some references how to classi/, classification of the data to specific disciplines. Um, yeah perhaps could be a better definition. And yeah. Yeah, OK,

mm hmm.

10. **I:** When you do research, do you create metadata for your data? Or do you describe your data in some way?

11. **B:** We have to, huh. Yeah, we have to, otherwise a lot of the information is lost. So, um what is perhaps important for us or the most important is we generate a lot of primary data, which include a lot of um digital objects like um data on microscopy images as well as the um derived data from this thing. And we absolutely need to introduce some sort of description of this, otherwise this information is quickly lost. This loses all the information content and the data lose the information content.

12. **I:** So how do you how do you go about it? How do you describe your data?

13. **B:** Right um. There are several ways. First thing is, we basically have to describe what our data, or how our data were collected, which subject they refer to. Um, this includes the, basically, the settings of an instrument which were used to obtain the data. This includes the description of the subject as well as treatments of the subject. And we use something, which are the metadata which either provided by the instruments, in our case very often microscope itself, which attach themselves to the primary data actually, so we have the metadata attached to the primary data. Or we have to just describe them separately. Usually as an entry in a lab book or an entry in a Word file and then attach them to the primary data. The same is true when we actually analyze data. So, we also have to have a protocol of how we analyze the data because otherwise it's also not reproducible. And to do that, we mostly do those by hand, but we also try to now introduce more electronic data management systems.

14. **B:** So, we have an institute-wide system to manage microscopy data, we have now the electronic lab notebook implemented only recently, still in the process of defining this. But yeah, I guess that about covers it. And, of course, we have the lab books which have to have links to the primary data all to be described the protocol for data analysis.

15. […]

16. **I:** So you talked a lot about the type of metadata that you create and that you write. What are the primary uses for this metadata, for these descriptions of the data? In other words, why do you do it?

17. **B:** Well, because if you just collect the primary data and they are not annotated properly, this information is lost basically. It's very it's very difficult to reuse it, if not impossible to reuse it. So, we absolutely need this later on. For example, when I have new students coming to the lab, right. So, we need not only the primary data, but we also want to analyze, to know what this data describe first and second, were they actually useful. So, we have to describe somehow their value. And this is also absolutely required now for publication, because yes, we use data when we try to support our claims, yeah? We use this as evidence to support our claims, but we also need to provide the metainformation about how exactly this data were collected. Where does the sample come from. A lot of publications now actually require you to provide this information together with the data. The same applies to, um to secondary data which we basically gain from the analysis of the primary data. How exactly they were handled. You need to describe in terms of metadata what exactly they do. Well controls, positive negative controls, all these things. So this is how we validate our methods and so on. Am I saying the right words or is this what you want, what you need?

18. […]

19. **I:** Could you please describe the role the metadata play when working in a team, maybe even internationally.

20. **B:** […] Yeah, it's mostly. Well I mean yeah, we do have collaborations but very often this case, we just share the data as well as the metadata. And as I said, very often they are just attached to all the primary data, so everyone could reuse this. Which is very nice and this sort of ensures the continuity and um the reproducibility of research. So we speak the same language with our collaborators but the same is true for working in the lab, right? So if we obtain certain results, yeah, we should be able to reproduce this and essentially we have to make sure that we do this in exactly the same conditions. So this is mostly a way to ensure the continuity of the data the completeness of course and the reproducibility. Yeah, I mean literally the way it's done is we can always look up what exactly was the setting of this instrument and you could always try to figure out using the lab book, how exactly the sample was prepared for collecting this data.

21. **I:** Right so you mentioned that it's quite a common practice that you would even reuse research data from others. What is your experience with the metadata of

research data from someone else?

22. **B:** Um, it's still not very widely spread. But there is a clear trend in um experimental molecular biology to make, to reuse the metadata. The protocols which were used and to attach them to the data. So the reason for that is exactly, because of the problems people have um faced after failing to reproduce the data so um let's say an good example. There are, as usual, there are good examples and bad examples in case of good examples you usually have well documented resources including data and the protocols which describe very nicely and exactly what to do.

23. **B:** And in this case you very often get similar results or reproducible results. There are not so good examples and this is not because people uh falsify anything or not because they did not provide sufficient metadata but probably because uh because some of the data or let's say the results already not the data the results, which are published, they are well filtered. Um there is this pressure for people to publish new stuff, so therefore you know a lot of high quality negative data never make it for publication and they are essentially lost. Thing there is now the increasing realization that they are as useful and as valuable and at least in my field, there are some attempts to actually establish online publication platform to also share this. They're not well used so far. Yeah, but, let's say that, yeah.

24. **I:** So have you encountered any problems when describing research data?

25. **B:** […] Yeah well, the thing is that properly describing and annotating data is um difficult, it's very time consuming. We also work in a field where it's not so easy to formalize all the treatments and our subjects, basically, with cells we work with. They are highly variable and then they mutate with time, so this is one of the troubles. And another trouble is that it's very difficult to figure out the correct amount of the description so of course, you want to describe this in as much detail as possible but at some point it becomes impractical. So, we have to find the middle ground between, you know, streamlining the work and also providing enough, let's say sufficient background information about the sample about sample collection and so on and so forth.

26. **B:** So yeah, we tried to figure this out. Another thing is that it's, I try to really instill this in the lab that everyone has to not only provide the data but also to describe them but it's not always the same (laughs) for people working in the lab

so. The simplest metadata is what the person writes in his or her lab book and the simplest and yet the most important challenge is that it's very hard sometimes to read someone's handwriting. Yes, it is, but this is true, right. So that's why some people start typing and then pasting this in but officially this is, you know, not everywhere accepted so. And the electronic lab notebooks are just taking off. So there could be some trouble with this. But we tried to find a reasonable, reasonable balance between streamlining actually the work and also describing what we do.

27. **I:** So do you apply any metadata standards?

28. **B:** No. I know that's, uh OK. Let's say some of them, right? So for example for microscopy data the metadata, because they simply attach themself to the images microscopy images are well-defined. So yes, we have some standards, which are actually used internationally, so. There are a bunch of software, which will be able to um, take up this metadata. Um but a lot of the manipulations are not well standardized. I know, I actually looked up, there are some um there were at least some attempts to standardize the description of the metadata but everyone runs into the same problem. So unlike, I know in medicine, where you could actually standardize the sample collection and the certain information, what we do, is very often based on research. So the questions come up as you go. So it might be quite difficult to become, but yeah. Uh so rather the answer to your question about standardized metadata collection is, it depends, huh, rather not for our everyday thing here.

29. **I:** Right, in what language do you actually describe your data?

30. **B:** English.

31. **I:** Any particular reason apart from that you speak English?

32. **B:** Um, no. Well in the lab I basically tell people that they could describe this either in English as an international language, which is spoken in science throughout. I mean even at home, though English is not my mother tongue and not for my wife, we actually when we speak about science we speak in English. Yeah, we switch to English, it's easier. But of course, because we are in Austria, it's also in German, right. So, people, who work here, they can describe this in German so um. But mostly English because I also after speaking to several of the students, they say that this is actually easier and this helps them structure their

narrative.

33. **I:** Right. So you mentioned already that you would sometimes use data from others, maybe even from another university. What's important to you, when it comes to the metadata, when you want to use research data from somebody else? What are the crucial things for you?

34. **B:** Good description of the experimental subject, how it was obtained. Um, this exact explanation of what the uh manipulations for the subject were. What exactly was done and how. Maybe also um. Ideally when people say, what they um, how they treat the data. So basically what are the indication of something which worked and which did not work. So it's sort of the criteria … for data filtering, which some people to subconsciously.

35. **B:** And this is unfortunately not very often obvious or clear. Because I, after I try to reproduce a thing and I write to people that we tried to do this way and it didn't work, they said, but of course because you used something different. This is something which is self-obvious for them but they did not really transfer this information to a publication or to a protocol. Um, yeah. The same also very strongly applies to data analysis, because this is again yet another sort of manipulation and for some people this is very obvious what they do and then they say oh, this is obviously not what we take into account, but this is not necessarily true especially if you try to do the automated analysis of the image so you actually have to say specifically, to state specifically the criteria for inclusion or exclusion of objects so. For the analysis. Yeah, I guess that’s it.

36. **I:** Imagine that you complete your current research project and another person wants to use your data. How would this be possible?

37. **B:** In our field, there are a few possibilities to share primary data. The one is called Figshare. This is an online resource and several journals actually now request the authors to upload their primary data onto this resource. This is what we're doing now. Um the problem there, of course is that yeah, you upload a subset of data but very often it's very difficult to upload the metadata. For example what exactly were we doing because that would necessitate copying of lab books and things like that. So, we provide some minimal information but if we, if there is a figure we provide the original primary non filtered data and we describe what we do with this data usually in the text and the materials and methods, people could at least try to do this. I know that people also now use

besides Figshare there is another resource.

38. **B:** I think it's not the ResearchGate but the ResearchGate is also possible to allow us to upload some primary data. And then there is Mendeley, Menedley data also could upload your primary data. And the publishers, some journals they basically ask you to provide the links to this. So in the recent publications, this is what we have done. Again we did not upload the whole lot because some of these are basically filtered because we have too much. So this is what we do.

39. **I:** I see. So are you aware of any research data management training or courses? Or even consulting.

40. **B:** Hmm, courses, consulting. No not really. I usually try to find this information myself. Yeah. That's the idea, you want to share your data as much as possible. We had an introduction to the [university institutional repository] but we found out that the system is not well designed for our purposes. Mostly because, you know, there is this possibility to share the primary research but um, you know, the moment you actually upload this, you could of course introduce certain embargo time but we don't want to upload everything so. Whereas the idea was that everything's actually uploaded as continuously and this is open for sharing. So this, I think was the biggest problem, we don't want to upload everything. Because there are a lot of negative data and so on. Negative in terms of low quality data. Um no, I can imagine that this would be something useful. But again, I think the biggest problem is that you know, like if he, if this has to be done properly, you actually need a person to do just that. To do the data management because this is not a small thing to do. So, and we try to find a middle ground. And optimize between really doing stuff and also, you know, making sure that the data retain the integrity.

41. **I:** I see so. Actually we're coming to my last question now. So imagine, if anything were possible really anything. What services or what support could the university offer in terms of creating metadata for your research data? What would really help you in that area?

42. **B:** Having a person who would actually really do that. Because as I said you know like we're, we're in the business of actually doing experiments and you know we always want to make it fast and sort of streamline the gain of knowledge, right? Um as much as data integrity is important, it really takes time. I mean, I know this personally myself, because if you want to do this properly it takes a lot of

time and a lot of effort. So, if anything would be possible, hmm like a position would not be um, actually, bad, to make sure that the data managed properly that everything is annotated and up to standard. That's, to make sure that the, basically the data retain their integrity and are evaluated from time to time so that you don't archive something which is not important. And really highlight something which could result in, you know, further knowledge gain. I think that would be good.

43. **B:** What would be nice, if the university would also pay for some of the systems, because they're not free.

44. **I:** You mean like the lab books?

45. **B:** For example like electronic lab notebooks. Yeah, I mean it's actually implemented on our institute level. So, but with, you know, as you upload more and more data you'll start to have to pay. You need to start paying for the just for the capacity of the. Again because this is what's needed, right. So yeah, I think the ideal thing would be to have a position for someone, who would actually do that, right. […]

46. **B:** I think it would be nice actually, if people um depending on where you work, would try to establish some sort of work groups on developing standards and establishing the sort of the good practices, the best practices in data collection, annotation and sharing but again, you know, like as I said. You know, what's required from us is more publications, so we need to continue doing experiments and you know um we all, we cannot spend a lot of time on just developing this thing ourself or you know organizing these groups or workforce. But I think that is not such a bad idea, where people would do this. I mean very often, I have to say that a lot of this thing is driven actually by the journals. So the good journals in our field, they actually request the authors to adhere to certain practices.

47. **B:** For example now in, it's a very specific example but it's actually a really nice one. Whenever we um submit a paper and we use some DNA constructs, we generate, a lot of good journals actually require you to provide the links to the public repository of these things which simplifies the material sharing enormously. I know that some journals actually now also request you, and this is a prerequisite for publication, to upload the primary data. And on one hand this is good because this forces the standards on the other hand you know like every time you are forced to do this like, oh man, now I have to go through all of, you

know, the previous five years of our data and make sure that everything is fine. So and, you know it's. I think the idea's that you do not do this retroactively but actually, you know, do it as you go and this is not always possible, so. Yeah a lot of those practices are driven by the journals and I think it's actually strong for us, because we want to publish. This is a good thing to actually make sure that data are coherent and reproducible. Yeah, so I think that's it.

48. **I:** So do you have anything anything to add? Maybe something that I didn't ask about? […]

49. **B:** I think what would be important too, is to introduce this concept quite early … Um when the students are educated, concept of data management. Have a course, which I teach in the university. It's mostly about writing, but I also try to introduce students to the concept of, you know, sharing data and at least mention a couple of things where primary data could be uploaded to. Because this is something what defines the scientific culture. I think right now with more and more data and essentially all of the data being digital. To make sure that what we produce is not just terabytes of unannotated garbage um, students need to realize, what they need to do with this in order to make, what they produce useful. So I think, an introductory course in data management with specific examples und good practices and, you know, basically description of what they need to do with their digital data, which is essentially everything, you know, is very important.

50. **B:** I mean I teach a practical course and I introduce how they need to keep their lab books written (unv.). And this is already quite difficult, because it takes time to develop the skills. I think for data management, this is even more true. That's why I think an introductory course would be useful. In particular things … how you manage the data, how you annotate them in a reasonable way, what exactly are metadata, which ones are useful for others. What are the standards and how they are stored, I think that might be good to define the future culture. Uh, I think that's it. Another thing which is very important, develop some tools, which make it easy (lacht) because this is very hard. I remember, I introduced the idea of the electronic lab notebook to one of my advisors long time ago when I was a postdoc and he told me straigt, point blank, that there is no such thing as electronic lab book and will never be. And … there were some systems but as all systems, you could make this either very complex and then it will take an enormous amount of time or you could make just the bare essentials and then it probably doesn't serve the purpose. So I think, we all need to define some

middle ground, where it will streamline the process. […]