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

This dataset is licensed under the Creative Commons Attribution 4.0 International License.

I: Interviewer

B: Befragter / Interviewee

(unv.) = unverständlich / unintelligible

1. **Wissenschaftler F – Teiltranskript, persönliches Interview (09.05.2019)**

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

3. **B:** My name is *Wissenschaftler F*. I work on the sha/ the morphology of plants, so their phenotype. I'm trained as a classical floral morphologist but I'm more, I work now on quantitative morphology of plants, mostly with flowers and seeds. My tool, the tool I use is X-ray computer tomography.

4. **I:** […] Could you please define research data in your discipline? What do you see as research data?

5. **B:** Our primary data is of course scan data, volume data. This is the primary data. So, what the machine produces a CT scan and produces. Our primary data is very heavy, so it's typically a few terabytes per project. Now after that we use this 3D data to um, we digitize it. So, we really boil it down to some important features that we like, that we want to quantify and this is more for metric data. Typically much much smaller I'd say in the order of kilobytes because it's very precise information. So that's the two forms of data we have and we run statistical analysis on the morphometric data, basically.

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

7. **B:** Reminds me of forms that one fills when one in a data repository, to be honest. So basically things that can be searched that help your data to be

searched. But maybe I'm wrong.

8. […]

9. **I:** If you would have to define metadata, the term metadata, how would you define it? […]

10. **B:** Yeah, I would define it as a descriptive file of a data, um, including a few entries that allow to perform searches on data entered by many different people.

11. **I:** OK, so do you create any type of metadata for your research data?

12. **B:** No. No, because I know, what I have. I don't need to search it. But when we when we put the data in repositories, we have to fill in such forms. Yeah.

13. **I:** Do you describe your data in any way. Not on the level of the data itself but some sort of a description, description file perhaps?

14. **B:** For myself?

15. **I:** Or for others perhaps.

16. **B:** For myself I just um I … The way I do it for myself usually is that I use the names that are um, I search my data because I have a lot and lot of data of course but I use names for my data that are descriptive enough so that if I search for the name I find what I want. So, I don't use an extra file. I usually make names that are informative enough for me to retrieve them, if I'm looking for them. So, the main constraint is, when the data was produced the data's yeah, when it was produced and its name. Usually they're both very informative. In case it's not enough yeah, I do use descriptive files I guess that would count as metadata, for scan data for example we always have a descriptive file, I guess that would count as metadata. And these descriptive files include, yeah. I always tried to keep this to a minimum because it's time consuming, so it has always to be kept to a minimum. Some people like classifying things and have a certain ease with this kind of administrative way of thinking. I do not. I tried to keep it to the absolute minimum. For example, as soon as I could, I stopped having a lab book. Some people continue using a lab book but I do not use a lab book.

17. **I:** […] So, is it common in your discipline to work collaboratively with others in

teams?

18. **B:** I am one of the most collaborative persons I know. I am currently working with more than 20 people across the world. From, I have collaborations with Oxford, Yale, German universities, Swiss universities, universities in Turkey, China, all around the world really. And yeah.

19. **I:** OK. So, when you work on a project collaboratively, as you do now. Could you describe the role metadata plays or the descriptions of your data, when you work in a team? Does it play any part?

20. **B:** Yeah, of co/ um. My role is to generate um three-dimensional data. I mean, I am. I developed the protocols to get good 3D data with X-ray tomography with flowers. This is what I did. And that's why I work with a lot of people. So, usually the people don't want to see the scan data. It's just cumbersome for them. They want the more informative version, the geometric, the morphometric data they want 'cause that's what they can process statistically. So, the scan data stays here. I don't need, of course, we have a way of describing it, usually small files and things like this but it stays here. They don't, very rarely people ask us the raw data because it is again gigabytes and gigabytes and it's cumbersome. As long as they know it's in our archive and freely available. That's enough.

21. **I:** OK. Do you ever use research data from others? Or have you ever?

22. **B:** I would say a priori no but I have to think. Research data from others, once maybe once for sure. I'd say at least once, yes.

23. **I:** OK, let's take this example. OK, how was your experience with the metadata for these records?

24. **B:** That was clean, it was clean. It was well done. I mean, metadata you really mean like um, a file describing, allowing us to use the data in a proper way, a descriptor kind of, right? OK, that's how I see it. That's how I see it. Yeah, it was clean and it was done professional.

25. **I:** So, you could use the data?

26. **B:** Yes, I could.

27. **I:** Was there any, were there any problems? Was there anything that you would

have wished, would have been done better perhaps or differently?

28. **B:** No, it was professional.

29. **I:** In which language do you describe your data?

30. **B:** English.

31. **I:** English. Are there any reasons for using English?

32. **B:** Um, for someone working in science of nature, so you know natural sciences, sorry, all the literature's in English. And so although our mother tongue may be a different, we tend to think in English at work and I certainly do. So, I write my, the language I'm the most comfortable with is French but at work I write and think in English and it is consistent. So, I read all the papers in English, I think in English, I write in English.

33. **I:** I see. Have you ever heard of metadata standards or do you apply any to your data?

34. **B:** We, when we have scan data, we do yeah well. Science, science requires reproducibility. So, when we scan something, when we do an experiment, we have to record a minimal number of parameters that would theoretically allow the experiment to be repeated. Therefore, we record these parameters which is about eight to ten parameters, almost always the same that we publish with our data every time. Because these are theoretically enough so that one could replicate the scan, the experiment. So we do, we have um basically a template that kind of arose not by itself but these are the parameters that I estimated to be needed in order to redo the scans at the same quality, same level as the ones that were done. So, it is 10 entries not more, 10 entries.

35. **I:** That's interesting. Is this common in the discipline, is that some sort of written or unwritten rule?

36. **B:** I think colleagues use something similar. I think I took something similar from a colleague and I just modified it minimally, yes.

37. **I:** OK, would you tell me something about these modifications?

38. **B:** Oh no, it was too long ago. I wrote this, the first really paper on this was in

2012 and we just kept the same format because it is enough to reproduce the experiment. And again, it's like ten parameters.

39. **I:** I see. OK, so imagine that you complete your perhaps current research project. And somebody else wants to access the data. How would this be possible?

40. **B:** If the data, if it's one of our published projects, the data is on [the institutional repository]. At the end of the paper, there's a link to [the institutional repository], where it is. That's it.

41. **I:** Do you always publish the data?

42. **B:** Yes, yes. As far as I know, yes, we always publish data.

43. **I:** What are the reasons for publishing the data?

44. **B:** Reproducibility. First some journals would ask for it. And also it's the way you do science nowadays. You have to be able to reproduce the results. And to get the same results, you would need well, the scanned data is useful let's say. Yeah, reproducibility, that's it.

45. **I:** You already spoke about [the institutional repository] a little bit and the metadata fields that you have to fill out there when you upload the data. But you said that you haven't done it yourself?

46. **B:** No, I haven't.

47. **I:** So you don't have any experience with it?

48. **B:** I'm very sorry. Our technician, she's here, if you want to ask her.

49. **I:** That's no problem. OK, so this is how it is. How do you do it? You work on the data, you publish your paper, then you want to publish the data in [the institutional repository]. but you don't do it yourself. Could you just walk me through the process?

50. **B:** When the paper's accepted, we upload the data to [the institutional repository] and we include a link in the manuscript and that's it. OK so we, what, our data is in a folder either on one of our servers or one of our hard drives. And when the project, the paper is accepted then we take the 3D data, upload it to [the institutional repository], get a link for it, put it on the paper, put it, include it

in the manuscript. And journals are happy because it theoretically allows to replicate the results, which is important. Which is of course much more stringent. Yeah, nothing.

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

52. **B:** Consulting? (lacht) Oh that sounds like something funny. No, I don't. I mean, I can imagine that if you're producing a lot of genomic data, then it becomes needed. I'd imagine, if you have a genomic-heavy group maybe you need to do this kind of stuff. But in our case, I think we're still on, yeah well hmm. We may change, we may need something more standardized in the future. As we are moving towards a field where we're producing more and more data faster and faster.

53. **B:** I mean our, the size of the data we produce in each project is growing actually, to be honest. And we may need to professionalize our way of doing things yeah, at some point. It will become needed to have a better organization soon, very soon um especially the throughput of our data production is going to increase possibly by a tenfold. So buying new equipment and I fear we may need actually some help, maybe yes. Let's see. In the future, I think that is something we may need at some point.

54. **I:** OK so how do you imagine this professionalization of research data management in your discipline?

55. **B:** No idea. No idea. I can imagine that a consulting, something like a consulting will be a lot of hot air and not too much real advice but I don't know, I really don't know. I have no idea, actually. I have no idea. Because one scan data is one gigabyte and we may be producing in the future ten a day. And uh I have the feeling that it's not gonna cut it without some kind of organization. My project and my idea is in the fu/. Oh am I talking too much or is that? So my idea in the fu/ is to produce, yeah … at least ten terabytes within a few months, three months and I think this will be about 10000, 20000 individual samples each of which about one gigabyte. And I think we're never gonna manage to handle that amount without some professional structure. I don't know. I have no idea how to do it actually. To be honest with you, I don't know. But this is the future. Very large datasets are the future. It's absolutely unavoidable. I mean the future, if one wants to be relevant in science today.

56. **I:** That actually leads me to my last question. We've already talked about it a little bit. But imagine now that if anything were possible, really anything. Finances don't play any role.

57. **B:** So, I'm back in [the country I come from], yeah, OK.

58. **I:** What services or support could the university offer you so as to meet your research data management needs?

59. **B:** Management or storage?

60. **I:** Everything that has to do with research data management.

61. **B:** Yeah, OK. Um oh, give me, give me a server with 50 terabytes on it yeah. I'd like much more space, much more space and I wouldn't mind, if you could provide some advice on a way of managing that. Yeah.

62. **I:** Anything else? We spoke about reproducibility and you said that metadata plays a part in that, finding the data that somebody put up online. Could you imagine any services to do with the metadata, with creating metadata for the data?

63. **B:** Yeah, I imagine it could be very useful. But just … it would have to stay streamlined. It should not be too time consuming, because time is really important. Um there's a cultural difference between me, I'm [from another European country], and Austrian people. Austrian people don't always value time that much, but time is really important and time is money. I know, that it's really important that things have to be done efficiently, not giving these huge file to fill in where there's not, there is just someone clicking, clicking, clicking for five minutes. You know what I mean? It has to be done with the thinking about the time of the person who is filling it up. I think this is really important. Because otherwise people will just go around it, see it like a hassle and just go around it.

64. **I:** So what do you think would help with that apart from making it easier?

65. **B:** I have no idea. I'm sorry, I don't have idea. This is not, I mean um. Something easily searchable, of course. Talking about my field really, in my field … The problem is, there is really a tradeoff between making it easier to search, easier to visualize and how much work we have to do, when we enter it. And I want us to have as little work to do as possible, really that's the thing. Maybe … If … okay,

okay yeah okay. Asking us how, because we have metadata files anyway. We have this, the parameters that are important to replicate the science. I think what would be useful, would be to ask us for our metadata files. Yes, to ask us. Because we have the files, we have something. And um if you make it that you can include our files, probably that would not be a bad idea. Because we have it already and it's probably going to be way more meaningful than something that is very general yeah. Something like that?

66. **I:** […] Is there anything else that you would like to say that has to do with the topic that I perhaps haven't asked you already?

67. **B:** We will need a lot of space in the future. More and more and more space. This is really going to be a problem if we don't have this space. We're going to need loads of space, tons of space, especially disciplines like mine. Because we can generate and we will. If, and this is not a bad thing, this is if everything goes fine and well, we will generate ten gigabytes a day. And at some point, this is going to have to be stored. And this is, if we are working really well. This is not like … if we can do that, we will publish rigidly well. It means everything is going perfectly. But we have huge needs of space, yeah. And we will have more and more needs for the space. It's not going to decrease and I don't know how to solve that but this is going to be a big problem.

68. […]