Module 3: how to write a README
Transcript

Below you find the transcript for the presentation "Presentation_Module3.pptx". This document is part of https://doi.org/10.4121/21399975, created by Cindy Quik and Luc Steinbuch, and licensed under CC-BY-NC 4.0

Slide	Transcript
1	Welcome to Module 3
2	As we just saw in the previous module, metadata is very important. Without metadata, the data is far less valuable and cannot be FAIR. In my opinion, metadata is most related to the F and the R of FAIR
3	When starting a new set at the data repository 4TU, one is requested to fill in such fields. This is an example of organised metadata, which, once in databases, provides a lot of opportunities for linking, indexing etc. A README file is something similar, a text that you can structure yourself. Interestingly, I didn't find any tools to transfer such metadata to a README file. In practice, I assume you can copy-and-paste quite a lot between such a metadata form and a README file, which is just one form of metadata.
4	I want to mention two current developments: standardizing of metadata and the use of vocabularies and ontologies & taxonomies. In the future, these might become important also for us as plain researchers. Metadata standards describes exactly which fields are available and also the used data format; the movement around vocabularies and ontologies is about using predefined terms and how those are related. On the long term, this offers possibilities for a really smart internet where concepts and databases are logically linked. Note that this information is mainly interesting for librarians, and those who create and maintain the fill in forms at repositories such as 4TU. However, we researchers should be aware of the developments.
5	In case you are interested: there are quite some metadata standards around on the Internet
6	As example of controlled vocabulary: here is one with geographic names, with for example the spelling of the Dutch city The Hague. I found some errors though: wromng names for The Hague, and: it is not the Dutch capital. It shows also some elements of taxonomy: The Hague ios a city in the Province of South Holland, which is part of The Netherlands etc. Also, it is logicaly linked to concepts "inhabitated place" and "provincial capital", which has elements from an ontology.
7	Let's go closer to the choices we have to make. An interesting one is how to name your dataset; a proper name also helps with findability, although search engines will of course scan also the keywords, description etc.
8	I took some examples from research dot wur dot nl. In my opinion, these are quite nice: the titles describe what the dataset contains, and it is also clear that it is a dataset at all.
9	Sometimes, the title of the dataset just repeats the title of the paper it is related to, or you see the title starting with something like "Data from..." in different forms.
10	With code projects, we see a similar diversity, and sometimes even very cryptic names. By the way, I looked into the "Software and scripts", for the author of this codeset, this were really too different things, although I could not figure out what was exactly the difference. Perhaps "software" are the more general

	functions, and “scripts” are for performing the workflow for a specific data analysis?
11	The advice from the WUR library is, in case you want to re-use the title of the connected publication, to add “Data from.. “ something at the start. Then it is indeed clear that it is a dataset, but in my personal view, you lose the valuable first words of your title, and in some overviews not much will be left.
12	I propose, to add – if needed at all – the terms “dataset”, “code” etc. at the end of the title. In an overview with for example different outputs for one person, you see directly which ones are dataset or code. But in an overview of a lot of datasets, to read directly information about the content. What do you think? And: in case of code, should we name the programming language, is that important at all? <Time for a short discussion. People might have different opinions about this>
13	Be aware that in case of software/code, there is often also a README file included, especially when starting up a project in a git-environment. Normally this is a README.md instead of a README.txt – .md from “Markdown”. This means a human readable text with basic formatting possibilities. The text version looks like this...(you see it is still very readable)
14	.. and the formatted version like this. By the way, also a README.pdf is allowed, if there are good reasons to do so. Other file formats should be avoided.
15	< handing out *Possible_topics_README* > Please have a look at the handout we created. Feel free to make notes on it. This is an overview of all possible topics you could put into a README, that we could think of. Which topics to pick exactly, and the order in which you mention those, is largely up to you. The topics in bold are the bare minimum, but in general: the more you can fill in, the better. Of course it should be readable and visually accessible, for example with clear headers. If it becomes a very large document, perhaps make separate documents, for example with so-called “codebooks” to explain in great detail. See as example also the README.md which comes with the handout << go to https://git.wur.nl/FAIR-data-for-ESG/creating-a-readme/, scroll down a bit >>. Because it is very short and it is in Markdown, I did not use headings but rather bold typeface. What do you think, is this readable, does this contain enough information? By the way: in my opinion, it often makes sense to start with a (draft) README or something similar right when you start with collecting data. It saves time later on, and might be useful for yourself as well.
16	Just let us have a look to this README. I hope you can read it. Does it contain enough information for you to re-use this data, would you call this a FAIR README?
17	Actually most of the information you are missing is in the standard metadata of the repository. In this case, the README is just meant to be additional to the standard information already available. Which is in itself a good thing, and saves time and duplication of information, but it is not a good README in the FAIR sense. If I download the data and README to my own computer, I have to search for this additional information later. A stand-alone README would be better.
18	Perhaps I talked enough for now! Short exercise: create a README for one of your existing projects or datasets, or create a template you can use for future projects. You can start from scratch, use our materials, or look around at the Internet. Note however, that when you are Googling for “How to write a

	README”, you will often end up at sources focussing strongly on README’s for code rather than for datasets. << organizing the hands-on part >>
19	<<Questions, discussion>>