What is data sharing?

Keith Russell, Australian Research Data Commons (ARDC)

Filmed at the Taylor & Francis Melbourne Scholarly Summit, 12 February 2019

The question of data sharing, what is it actually? It's that practice of making the data used for scholarly research available. And we see data very broadly. It's not just those tables, it's not just that database. It is all that underlying material that was used to create the research and, by making that available, you can actually reproduce that research. And that can be a huge array of different materials depending very much on the discipline and very much on the research conducted. So that could be text, that could be images, that could be video, that could be spreadsheets, that could be transcriptions, that can be audio recordings, could be video recordings, or can be drawings, can be notes captured in lab notebooks, and it can also just be tables of data or databases. So there's a huge array of different types of materials that could be considered data. And from our perspective, it's very much, "What are the materials that are actually useful to provide so that others can reproduce the research findings and understand how the research findings were created and build on those?".

So, we're looking at the motivations in this study. Our top motivation was increasing impact in visibility of their research. That was a strong argument for researchers to share their data. Other arguments were, the public benefit of making it available so that others could use it and build on it, increasing transparency and reuse of their research, and getting proper credit for the data by having a citation available over that data.

So, if you're looking at researchers and their drivers, what we're also seeing is drivers from outside, from other organizations, encouraging researchers to share their data. One of the drivers there is funders. Funders are starting to slowly but surely push harder and harder on to researchers saying, "Well if we funded research, we want to see what you produce and where the money went." So, one strong example is Wellcome Trust, who's taken a big lead on this and actually pushed for open, making the data available. But you're also seeing that from the EC as they're thinking about it and here in Australia, the ARC and NHMRC, at this point, strongly encourage the sharing of data and making the data available.

Publishers are putting up data sharing policies, including Taylor & Francis, a number of others are definitely also pushing in that direction. Government in different countries has set up open-data initiatives. In the US, here in Australia there has been a number of works in that space, and in Europe too.

So a few considerations in STEM, we're seeing as we are engaging with a wide array of disciplines within STEM. This image here is from bioinformatics, and it gives you an idea of the scientific workflow and the process of how they treat data and how they share data, and where they treat and share data. As you're looking at STEM and as you're seeing developments within STEM, we're seeing a lot of push, first of all for more and more larger machines that are collecting more and more data, new software that's enabling much more analysis over that data and the generation of data, and that's raising questions about, "Well how can we possibly manage all that data?". As huge volumes of data being spit out, how do you select what you keep, what you retain, how do you tie that together?

There's also a whole bunch of opportunities in this space. And this comes to that point about machine learning and artificial intelligence. As we are creating more and more data out there, there's huge

opportunities in bringing data together and combining data to provide novel research and novel findings. But that means that the data that's being produced has to be able to be tied together, and I'll come back to that when we get to the FAIR data principles. But making sure that data can be read by machines and humans, and be transferred backwards and forwards, is an interesting challenge, and it's one of the things that also offers a lot of opportunities.

Other developments in this space, in STEM too, increase the collaboration across disciplines. So that means that data being produced in one discipline and another discipline can actually be combined and can be brought together. And that's providing novel findings but that also brings challenges with it.

One of the questions we frequently get is, "What does data really mean in humanities and social sciences? Do we have data?". A starting point is the idea that data has to be tabular or it has to be in a database. And we'd definitely argue, "No, that's not the case." Data can be all sorts of materials and assets. So we see that in a broader perspective.

The scientific workflows that you see in STEM don't usually apply that directly to humanities and social sciences, and often that's actually part of the creativity within the humanities and social sciences, is turning that around and thinking in different ways and using different routes to come to findings. You are seeing certainly in the digital humanities, quite a strong push towards open data principles and using data visualizations to make that data available and visualize it in a meaningful way. Humanities data is not always easy, it's quite complex to untangle, and it's not always meant to spark joy. There's sometimes complexity and difficulty within the data.

Below, you'll find a number of different tools that are out there, different platforms that are available for different types of humanities and social science data. But because there's such a wide array of disciplines there, there's also a wide array of different platforms and tools out there.