

## Module 2 Video 1: How to Understand the Data You're Working With

[00:00:00] Hi, it's Heather. And it's in this video, we're going to talk more about how to embed equity and ethics into your data journalism project. And we're going to talk about one of the key tools, which is something that we called data biographies.

[00:00:17] So data biographies are part of the data collection and data sourcing, ethics and equity for data journalism framework and data biography's are where you spend time getting to know your data and your data source as well as you would any other source that you would put in a journalism piece. So it does not matter if you trust the person that you get the data from. It does not matter if you trust the organization that you got the data from. Oftentimes I hear well, if you're getting data from an organization that you trust, you can use that data after many, many years of experience. I'm here to tell you that is not a good plan, not because you shouldn't trust an organization, but because the assumptions that are embedded into that organization's data may not be made explicit to you by that organization. Not intentionally hidden. But maybe just they haven't engaged in the best practices yet of how to communicate data to journalists. Every time, every single time you put data into a data story. Every time we use data in journalism, it is essential that we ask whose worldview is embedded in this data and how is it affecting the data?

[00:01:40] Sometimes people come to me and say, I'd like to only use objective data. And I say, that's not possible. You can use is transparent data.

[00:01:51] Every data, as we've seen throughout this course, and we'll continue to see through the rest of the modules. Every data is embedded with someone's world view. All seven steps embed a world view into the data. It's unavoidable. The only thing you can do is to be very transparent. And when you're using data that you've acquired from a source, whether it's somebody that you trust a lot or somebody that you don't know at all, you need to take the time to really get to know exactly whose world view is embedded into this data. And the best way we know how to do that is a data biography. And I'll include a link in the resources to a data biography template. Data biographies are not complicated, but they can be a bit time consuming to construct.

[00:02:40] So if you're going to use data, you need to know who collected the data. What was asked or measured in that data? How the data was collected? Where the data was collected and where the data was stored? Why the data was collected and when the data was collected? This seems basic and obvious, but you would be surprised how much data does not come accompanied by this. And if you don't know the answers to these six questions, you should not trust or include the data in your data story.

[00:03:16] We're gonna look at a quick example from a very excellent organization, the United Nations. Progress of the World's Women, very well-funded, very good reputation organization. And if I'm doing a report on violence against women and I download their data set, which I would trust because I do trust the United Nations, this is the data that I would download from their website. And this is the average violence against women over the course of a lifetime for five countries over 10 years. So we have data from Rwanda, Malawi, Sweden, the United Kingdom and Australia. And you can see that the data shows that there's quite a bit of movement in the prevalence of violence against women in these countries and that the rates are going up and down quite a lot over the course of 10 years.

[00:04:16] So if I want to put this in my data story, I need to take a minute and think about this. I personally happen to have done a lot of work in the violence against women field and a lot of research. So I would be very skeptical about whether or not the rates are actually going up and down at this very fast, trending rate? I mean, these are very big changes in lifetime rates of violence against women within countries in a very short period of time.

[00:04:48] The way that I would check and verify whether I wanted to include this data in a journalism piece I was doing about violence against women would be to create a data biography. This is where you have to go digging. Now, the United Nations, one of the reasons the United Nations is a very good organization. You can't trust is that they do usually provide you with the information that you need to really dig deep and you need to look in the data codebooks. You need to look in the metadata. You need to call the people that have provided the data and call the

people who have provided the data. And all of this sounds like it's going to take a lot of time. And in fact, it is going to take a lot of time. And it is probably one of the most important investments in terms of embedding equity and ethics into a data story. If you're using numbers that you don't know, the details of a data biography for the chances of you getting it wrong or saying something that's unethical or or biased towards or against a certain group of people are very high.

[00:05:59] So, for example, if we look at just Sweden, we can through our data biography, we can see that these numbers are being collected from very different groups of people. So that answers our who question in our day biography of the who, what, where, when, why. The who. The first number is that that data is being collected by women between the ages of 18 and 64 who were ever married. And the most recent number from Sweden is being collected from women who are ages 18 to 74 with no specific relationship requirements. So the difference between those two numbers does not actually represent a change in the rate of violence against women. It represents a change in who the data was collected from. So this is how easy it is to accidentally put misleading, unethical data into a data story.

[00:07:00] Same thing for Malawi. We can see that. Who the data was collected from. Really, really changed. And it's not just who. I'm using the WHO example on these two charts, because it's probably the easiest to understand in a very short amount of time. But the other questions are also extremely important. How the data was collected is extremely influential over the results that the data is going to show. What methods were behind the data collection, design and process? Were the data collected in person where they collected automatically? Were they scraped? Were they cleaned? So, in other words, were kind of unusual data points removed, which is often a very bad idea. Where was the data collected? Where was the data stored? Why? For what purpose was the data collected? Sometimes we see very interesting data sets that are collected for the purpose of testing out a new methodology in data collection. So if that data had that has a new data collection process tells you something very different. You want to check into the why before you just use it in a data journalism story. So building a data biography is probably one of the most essential, straightforward steps to embedding equity and ethics into your data journalism process.