**Module 9 – Dave Tarrant – Choosing the right format for open data**

**Choosing the right format for open data**

I’m Dave Tarrant, I’m the Trainer and Data Scientist here at the Open Data institute.

**Why is the format of open data important?**

So the format of open data refers to how that data is made available and understandable for both humans and machines, and as well as considering the former, it’s also critical to consider the structure of the data, and also how it is delivered to the users.

**How do you choose the right open data format?**

So choosing the right format is critical to maintain the usability and the access, and there are three main characteristics, you choose the format, the structure and its delivery mechanism. So when choosing the format, the most reusable format is likely to be that in which it is already managed and used. This will be closest to that context in how it’s used currently, but this format might not be the most widely available, widely understood, so it might be necessary to publish in a format like .csv in addition to that other format. So as you can see, you should also consider publishing in a number of different formats, and also be prepared for any particular format’s obsolescence.

When it comes to the structure, there are several different structures that data can exist in. So the most common one again is tabular, where you’ve got things in rows and columns, this is used for things like expenses data. But there’s also hierarchical data which represents relationships between people, so think of your family tree, and there is also network data which is very hard to think of in terms of a file format, but this represents the links such as those between your friends on Facebook, they can go in any direction.

And finally you need to consider the delivery mechanism, so do you make it available as a file download? And if you do, you should carefully consider the size, so that it’s easy to download quickly and it’s not going to take a long time, the frequency of which it’s updated, so does it need to be updated regularly, monthly, and also the terminology that you use, to make that understandable. You should also consider if the data is more live, like transport data, that it might be more wise to provide a machine readable feed of that data and that would be more appropriate than a file based delivery.

**How have formats changed in recent years?**

So we encourage the broadest range of formats, structures and delivery mechanisms for data to be used. But core to all of these is the .csv format, the most commonly used open data format online. Beyond .csv when deciding other format structures and delivery mechanisms, it is critical to engage users and for users to engage providers of data. This saves both parties a lot of time.