**Module 12 – Dave Tarrant – Excavating hidden data**

**Excavating hidden data**

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

**What is ‘hidden data’?**

Open data is not always easy to find. We as humans consume things visually, and so we looking at pretty webpages and images on these webpages that cue us to where we can get data from and download it. Computers don’t operate in this way, they like structure, logic, clarity and mainly text. So the challenge for us is to understand how a machine likes to consume data and be able to get into its mind-set to be able to excavate this hidden data.

**Where is ‘hidden data’?**

The web is a data rich environment, and many websites already bring together data from various different sources in order to present you with a collated set of information.

So think of planning a holiday. Many holiday websites already bring together a rich amount of location data, pictures, reviews, prices and maybe even transport data that allows you to plan your route to get to a certain place. So what we want to be able to do is realise that we can exploit these same data connections that exist between the websites, so that we can get access to the data ourselves to present it in a whole new context - to build a new set of information and thus a new set of knowledge.

**How do we access ‘hidden data’?**

Some websites give you access to data simply by adding and extension to the end of the page you are on. So if you can see that there may be some sort of data on the page, you might be able to add like a .csv to the end of the page and it might download you some data. One of the most common formats online is actually the .JSON format, and you can get access to all of the data off of gov.uk by adding a .json to the end.

Other websites might give you feeds of data, such as news websites, and we’re used to consuming these within mobile applications. Beyond these it gets a little more tricky, some data you may have to look for from the source code of the page, where it’s been embedded, these websites are already using it, so you may as well look for it within the page itself.

Finally, many sources of data give you a specific way to access it, a machine interface or an API, so you might have to learn how to use one of these.

**What is the benefit of using this ‘hidden data’?**

So by using data that’s already an integral part of an existing service, you can guarantee a higher quality of relevance and usability. Additionally, you’re accessing it in the context in which it’s already used, rather than through a separate publication platform.