Challenges to Data Integration

So far in this course, we have discussed different types of data including genomic imaging, biological and lifestyle data. Modelling and analyzing such data, allows us to investigate different facets of human health. However, it is by integrating this data, that we can deal with the complexity of different conditions, and gain a comprehensive understanding of human health. Allowing us to characterize individuals much more precisely. Precision medicine will not reach its full potential, unless we effectively share data on a global scale. There are three main challenges to global data sharing. First, biomedical and healthcare datasets sit in silos. They typically reside within the IT systems of a particular organization such as a hospital, and they're stored in relational databases which are not designed for data sharing. Second, linking entities between different datasets is not a trivial task. In Scotland, patients are identified using a ten-digit number known as CHI number, allowing their healthcare to be linked across time and location. However, many countries do not use such unique identifiers, and even if they did, how about identification on a global scale? Third, when sharing data worldwide, there is the issue of multilingual content, and even when using a single language, there is sometimes ambiguity as to what particular terms mean. And people often use different terms to refer to the same concept. In the next few videos, we're going to have a look at different approaches to these challenges. We'll be introduced to the graph data model, which effectively addresses the first challenge, as it allows for the simple and flexible representation of data that can be easily merged. Our focus will be on the RDF data model, which uses global web identifiers. Thus addressing the second challenge. Finally, we'll familiarize ourselves with ontologies, which allow us to standardize terminology, providing a solution to the third challenge discussed. So let's get started.