Video Transcript Lesson 2.2 What is Linked Data

This is a quick introduction to the concept of linked data. It's a fairly non-technical introduction, so even if you don't know anything about web programming or how the web works, this should still be fairly useful to you. So you may have heard about linked data recently in the news, because Google and Facebook are starting to push it out into their products.

So the first question, I guess, is what is linked data? What is it used for, and how do we produce it? Now, if we go back and we think about data by itself, there are many different types of data that we use today. There are images, there are things like Excel spreadsheets, there are videos of a variety of different things happening all over the world. There are websites that include pictures and text and links to other websites.

So there's a lot of data that's accessible to us throughout our lifetimes. And one of the really cool things about the web is it put that data out there so that it's really easy for us to get access to it. So the web allows us to mix pictures with text, with charts.

But not only that, it allows us to link from one document to another. And using these links, we can follow them from one document to the other and figure out a bit more about the subject that we're reading about at the time. So there's this really cool discoverability aspect that the web unlocked between all the different types of data that we have.

So this is really a great thing for humans, but computers are still kind of back in the dark ages. They don't understand what's on these web pages. They understand that this is an image, but they don't understand what the image is depicting, when it was taken, any of that information.

They understand that this is a link, but they don't know the relationship that this link has to this web page. They don't know if it's a random link that's going to take you to a spam site, and they don't know if it's actually a useful link that's going to take you to, let's say, another Wikipedia article about it. So the big problem that computers have with the data today is that they don't understand it, truly understand it, in the way that humans understand the same information on a web page.

So what web developers tend to do, and they spend a lot of their time doing, is taking these web pages and deconstructing them. They take the little bits and bytes out of the page and package them up in a way that is understandable to computers. So we do things as software programmers, as web developers, that take the pieces of data on the page and package them up in such a way that express things like names, and birthdays, and moods, and locations, just so that computers can understand what we're talking about.

Now, there are two problems when it comes to link data on the web. The first one is a common format problem. It's a packaging problem.

What's the best way to express this data on the web? Today, we have HTML, and JSON, and XML, and CS comma separated values, CSV, and RDFA. So there are a variety of different ways of expressing data on the web. And then the other problem is, how do we link all of this data together? What do we use to link these pieces of disparate information together on the web, such that machines can understand it? So the easiest way to represent data to computers is in this property value mechanism.

So this kind of looks like an Excel spreadsheet, because that's kind of the simplest way of representing the information to a computer. So we have properties and values. So this name property has a value of Frank.

This birthday property has a value of January 3, 1985. This mood property has a value of happy. So this kind of creates this island of information that the computer can kind of understand.

It knows how to process it, because it has a pretty regular format. The problem is that it's not linked to anything. It's kind of out on its own.

And there's no way to figure out what Frank's relationship is to the rest of the world. So that's where linking comes in. If you think back to a website, websites have links to other websites.

We need the same kind of thing for our data. So we have this description of Frank, this data. And in it, there's a piece of information that's actually a link to other data.

So now we know that Frank is linked to Jan in some way. And if we look at the property, it's the nose property. Frank knows Jan.

And that's a pretty powerful thing for a computer. The computer can now use that information to figure out that there's a relationship between those two. Another way of representing this information is not necessarily an Excel spreadsheet, but this kind of structure called a graph.

Now, Google has their knowledge graph, and Facebook has their open graph. But at the most fundamental level, they're the same thing. It's a graph of information where each node is kind of a subject, something that you're talking about, so like Frank or Jan or Tim.

And each subject has information hanging off of it. And each subject can also be related to one another. So Frank is linked to Jan, and Jan is linked to Tim.

The way Frank is linked to Jan is through a nose relationship. And the way Jan is linked to Tim is through a parent relationship. So if you were to ask a computer, who is Jan's parent? It would go and find Jan, and it would follow the parent relationship and say, Jan's parent is Tim.

And that's how this graph of information gets built. Now, this is all linked data. It's all linked together.

And the really cool thing is that these green lines represent the borders of websites. So this is one website, this is another website, and this is yet another website. So linked data is data expressed on a website that can traverse via links to other websites.

And that means that the web becomes this global information repository where you can ask the internet and the web basically questions like, who is Jan's parent? And it would be able to start on one website, follow the link to another website, and answer the question for you. Now, we have this kind of cool thing on the web, this universal ID mechanism called a URL. And computers really like to be very specific when you're talking about things.

So if we go back to this graph, computers are going to have a hard time with this information because we're not specific enough. We're not specific on which Frank we're talking about. We're not specific about which Jan we're talking about, or even the words and the terminology that we're using.

So what we end up needing to do, at least with computers, is we end up needing to use URLs to identify things. Now, this works really well for us as well. So the main thing that you use to go to a website is a URL to the website.

If you want to share a link with a friend, you send them the URL. And you know that when you send them that URL, it's a universal identifier. And when they put it into their web browser, they're going to see the exact same thing that you're seeing.

So we want to use URLs for everything in linked data, basically. So to a computer, this is what linked data looks like. There's a URL, which identifies Frank.

There's a URL that specifies a knows relationship. And there's another URL that specifies Jan. Now, don't be afraid.

You're never going to see the information in this way. But for a computer, it's really useful, because now it knows that it can start here, and follow this link, and end up here, and find out more about who Frank knows. This is really how the global knowledge graph gets built.

Piece by piece, every single website publishes their data. All the data links to one another. And search engines and web crawlers can then use that information to answer interesting questions, like who is the President of the United States? Or how many people live in Libya? Or find all of the local offices that are up for re-election this year.

So there's a lot of information out there that's floating out there that's not in linked data form. And if we put it in linked data form, computers are going to be able to use it to make answering these questions much easier. And two examples of that today is Google and their knowledge graph.

If you do searches in Google, they now utilize linked data to answer questions directly. You don't have to go to a website to figure out what the answer to the question is. You just ask it directly, and it'll answer it for you.

For example, converting feet to meters. Or if you want to figure out what the President of the United States' birthday is, you can ask that. You can type that directly in, and it'll tell you what it is in the search result.

The other company that's really using linked data is Facebook through their Open Graph protocol. So Facebook asks people that create websites to tag information in the web pages so that computers can understand it. They use a technology called RDFA to do this.

But basically, what that does is it highlights pieces of the page that are important to a computer so that the computer can understand what the page is about. So linked data is not this kind of far-off concept. It's something that's being used today.

Now, if you want to find out more about linked data, there are two websites that are pretty good that you can go to. The first one is jason-ld.org. So this is a format for linked data on the web. And the other one is rdfa.info. Both of these websites have resources for finding out a bit more about linked data and how it's used and created on the web.

This entire video is released under a Creative Commons attribution share-like license. That means that you can copy it and share it with your friends and remix it as much as you'd like without asking permission. If you have any questions on this tutorial, please contact me.

This is my Twitter handle. And I'm also on Google+, under Amano Sporny. I hope you enjoyed this introduction to linked data.

And if you would like to, there are some follow-up tutorials on JSON-LD and RDFA as well. All you need to do is search YouTube or Google, and they should pop up.