h2lod

Michal Sokol

May 8, 2012

Contents

Preamble

sjezpz:192034 - but http://sajrmajaz.zlzpp.com/zedze/192034 really - or, as its friends like to call it, 192034, was a happy yet curious datum. Like a lot of data, it liked to be thought about as a resource; "I am resourceful" he was heard saying a lot. 192034 had a lot of humour, too. He was famous around his entourage as a witful pun maker. 192033, 192034's brother kept on complaining that "He isn't that funny."

Funny or not, resourceful for sure, 192034 was a whimsical being. It attracted other data like a magnet. Was it due to his charisma? He doubtlessly had an aura. Data would come and link themselve to 192034. His circle of aquaintance became larger and larger. Some might call him a leader. "I don't put sugar in my coffee," he liked to say. He was a datum that liked things for what they were. For him, the cup of coffee was always half-full. An optimist humanist, but instead of human, data. He realized that data are more than just values; he had nothing against couples, but believed in triples instead. Like most knowledgable datum, 192034 owned an owl. A beautiful one. "It gives me meaning," he said. When he took it to the owl-er coaster, he "could see SPARQL" in its eyes.

Yet, this guide isn't about 192034. Not even about his brother, 192034. It isn't either about their good friends 192011, 192289 and 19290, although it would be possible to write a good deal about the latter and its constant discontented manners. This guide isn't either about their direct referencial—home—sjezpz. The reach of this guide aims to go beyond what's visible to a simple datum; no matter how symathetic we feel for them as individual, we must study them as a whole. As women and men of science, we shall put our emotions aside for the greater good. Because places like sjezpz, there are many. All these agglomerations of data—dataset, really—are interconnected. They are a part of a larger picture; a perpetually expanding galaxy, the Web Of Data.

Chapter 1

An introductory Introduction

How about some coffee to start with? Some music in the background, maybe? This really doesnt have to be so formal. Yes, this is much better. You dont mind if I also make some coffee while talking with you, do you?

There is such a rich vernacular when it comes down to describing coffees. People would use various words like "fruity", "charismatic"; my favorite being the former. Same in enology. Where did people find all these words? Is it possible for a coffee to be "spicy"?

George Orwell was a great author. He wrote a good deal about languages. One thing he wrote was that:

> Language can also corrupt thought.

I would have never described a coffee as "spicy", hadn't I heard that depiction before.

Orwellian vocabulary put aside, coffee by itself also has a strong influence on our brain. There is a rumor saying that Voltaire[1] downed fifty cups *a day.*

Coffee is a lot like Linked Data. There are beans. They come from different places. They each have their own caracteristics. Some are hard to find; some you can't really trust. They are raw. You can't eat it like that—I wouldn't try it. You can't use just one; they are truly useful in a handful.

Then depending on what you want to do with your coffee beans, you need to process them. Espresso? French Press? You need to mix the beans together and grind them. Then only can you truly appreciate the scent.

[1] When he was writing *Candide.* He had the reputation of being a huge coffee drinker. He also wrote an interesting essay on it.

Figure 1.1: Coffee beans are like data; they are held into silos. Wikipedia is a good example of a data silo.

See, now it really is "fruity". It's only after mixing and grinding the beans together that we realize the coffee's real aromas.

Once we add the water, all the work has already been done. It's the final step. After that, the coffee is ready to be served—sipped—while it's hot.

Figure 1.2: It's only after taking raw resources and mixing and grinding them together than we get the most of them—in this case, the coffee I'm sipping right now.

Let's take a step back; shall we? To summerize, making good coffee requires to:

1. Find the beans (resources) you need
2. Mix these resources together
3. Grind them; transform them; serve it hot. In a nice cup. And share it with your friends.

That's the secret Voltaire knew while he was sipping his 37th cup of the day at *Cafe Procope* in Paris. He knew—without knowing it—what the semantic web was going to be.

This guide's structure is roughly representative of the steps described above. To begin with, we'll talk about data. Then we'll look at how they are stored; where they live. Finally, we will see how it's possible to brew our own "coffee" with it.

Before that, let's define a few terms.

1.1 Terminology

Semantic web, Linked Open Data, Linked Data, Open Data; there are a lot of different terms that need some clarification. Beware—some of these definitions aren't set in stone; they evolve, they change.

Semantic Web is a set of technologies aimed to structure data on the web to make them interoperable and understandable by machines.

Linked Data represents data published according to a guideline that emphasizes the relationships between each datum.

Open Data are data that don't necessarily follow *Linked Data*'s guidelines. They are open, thus available for anybody.

Linked Open Data is *Open Data* that follows *Linked Data*'s guideline.

Chapter 2

Meeting Data: A trip to LOD's capital

Data are the basic inhabitants of the universe described in this guide. As such, we must take time to understand and study them. Better than that; let's dive into applied anthropology and imagine that we ourselves are data. Yes! What would be better than a complete immersion after all?

2.1 Destination: DBpedia

So pack your stuff and get ready because we are leaving soon for our field trip. Don't forget to check-in your ticket online if you don't want those really bad seats. We'll be departing to the *happening* place; the actual *capital* of the Linked Open Data continent, *DBpedia*.

Geographically speaking, *DBpedia* is at the center of the *Linked Open Data Cloud*[1]. It is bordered by *Freebase* on the west, *Gutemberg Project* on the east. Its population consists of over a *billion* triples[2], and it's growing everyday. It is safe to say that *DBpedia* is a really cosmopolitan place: data from all around can be found there, in a vast diversity of languages.

Don't forget to take your coat, just in case it's cold. Well, I think we are now good to go. Whenever you are ready, go to this URL[3]. I'll meet you there.

[1] You can look at it here: http://richard.cyganiak.de/2007/10/lod/

[2] As you will soon discover, data don't believe in couples, but in triples.

[3] http://wiki.dbpedia.org/About

2.2 The customs: Your foaf identity

After what seemed like a terribly long trip—accessing a URL is way longer than it seems—you finally arrive in LOD's capital, DBPedia. Looking around, you see a long line in front of you. The officer asks people to prepare their foaf[4] *file. You see a booth with a bunch of forms on it.*

Family Name	_____
Given Name	_____
Email address	_____
Age	_____

Table 2.1: LOD Custom form

In the semantic web, data need to be structured in order to be processed by the computer. In order to publish information—such as your identity to that custom officer who is starting to get impatient—it is required to use a common *vocabulary*. Just like when you communicate with someone, you both need to speak the same language.

Foaf is a vocabulary[5] that describes individuals and their relations. If you take a look at the project's homepage[6] you will find a list of all the attributes you can describe—such as a person's email. Using such vocabularies to describe data allows it to be processed automatically, which is what semantic web is all about.

What is interesting here is those pieces of information you provide, your age for example. Those numbers; strings; images; they add up to define who you are. They are the basic entities in the *Semantic Web*; the "Data" of the "Linked Data". They are raw. They only really make sense beacause we used a vocabulary—*foaf*, in this case.

Data are elementary descriptions of things. They only make sens in a context. They are the base upon which information and knowledge are built.

By giving a context to a datum, it becomes an *information*. Information is the basis upon which *knowledge* is built.

[4]foaf stands for Friend of a Friend. It is an *ontology*, which will be described later on.
[5]You shall see many of those through your journey in the LOD.
[6]http://www.foaf-project.org/

In order to be understood by a machine, this knowledge has to be formally written. Just like when you need to fill-out a form for the customs. The 2.1 form could be one of these forms.

In later parts, we will see the formalism introduced by the semantic web standards.

After filling the form with your best possible handwriting, you proceed to go back to the line. Things went well, you didn't have to wait too long to get your lugguage. You got your stamp: you are officially a resource. All of this is getting exhilarating.

2.3 On being a resource

Applied anthropology, that's the nature of this field trip. The web is full of data, and these data are there to describe *resources*. Resources are the *context*, and data are the attributes. The link between resources and data provides *information*; data aren't raw anymore.

Resource can be anything that has an identity. Resources can also represent abstractions.

In the Semantic Web, resource is a very important concept. A resource is a "thing". Your computer's screen is a resource. This box of chewing-gum is a resource, too. You are *yourself* considered a resource in the Semantic Web's terminology.

A resource is like an "instance"; it has a unique identity.

The *world wide web* employs URLs to identify documents. The Semantic Web uses *URIs*[7], to refer to resources. A URI has the same form than a URL, except that it can refer to anything. It isn't limited to documents. My turtle could have its own URI if I had a turtle.

It is the unique identifier that represents something. An alternative name to *Open Data* could have been *Open Resources*.

Here are some example of resources URI:

Since resources are valide URLs, they can be dereferenced. If you copy and past the above URIs in your browser, you will see information on that resource; you would see the data bound to it.

[7]Universal Resource Identifier

This forms the basic idea of the *Linked Data* structure. Resources are agglomerations of data. They are dereferencable. Moreover, the format of the data we get by dereferencing a URI is *machine-readable*. If that data uses a vocabulary you are familiar with, i.e. *foaf*, then you know its structure. You can extract information you want from it; much like you would extract information from a database you know the schema of.

What it comes down to is that:

Semantic Web Structured agglomerations of data.

Or even more concise:

Semantic Web Structured resources.

Structured data is the goal of the semantic web[8]. It's by structuring the resources that we build knowledge.

On a more practical note, chosing a URI for a resource is arbitrary. Chosing a URI for *you* would be done using a domain you own. Dereferencing it would return the *foaf data* that describes "who you are".

If my domain is, say, http://sajrmajaz.zlzpp.com/; I could refere to my turtle—the one I don't really have— as http://sajrmajaz.zlzpp.com/shrodinger.

It is the head full of new concepts and ideas that you decide it's time to rest. You've had a long day since your journey started. Tired from the long trip, all this new world fascinates you. You want to learn more, to find out about all these small details such as what's the cuisine *like in* DBPedia. *You find it hard to fall asleep, but when you do, all the tiredness, the jetlag, it's all gone. You are all rested; ready to start your second day in the capital of* Linked Data.

[8]It will be discussed later

Chapter 3

Interacting with the locals

Resources love to talk. The kind of discussion one has with a resource is often deep and enlightning. Somehow they know how to convey *information*. It's not mere chatter. They go straight to the point. They avoid saying anything useless, they employ no superlative. Far from using *pathos* or *ethos*, they stick with *logos*. You can't have drama. They won't complain either. Their discourse remains neutral. Just like Howard Roark, they stay objective.

The truth is, resources don't rely on emotion. No matter how frightening the question you ask them is, they won't disclose any apparent surprise. Some say that resources are cold, lifeless beings.

But behind that factual-minded agglomerate of data, there is just a simple lad doing his job. Not unlike us, he also needs to change his mind comes the week-end. Not unlike us, he needs to socialize, he needs to take the facts out and let the burbon flow in.

Some of us go to bars. Resources go to *endpoints*. They all hang out there. Getting to know each other and bound.

This is why today we are going to head to such a place to have our first encounter with a resource. The place we are about to visit is called *DBPedia's Endpoint*. Needless to say, it's very busy and crowded. But the experience will be worthy.

Before doing that, there's a small thing we need to discuss: language. Well, yeah. If we are going to communicate with resources, we might aswell learn their language, right?

3.1 RDF: The language barrier

English and other natural languages are wonderful framework to think, describe and talk. But resources don't need to do much more than describe things. They might describe who a poet is; they won't be writing poems themselves. For this reason, the language they speak is fairly easy to understand.

RDF the *Resource Description Framework* is W3C's abstract model for describing resources.

RDF is semantic web's way on conveying information. More than a language, it is a framework—an abstract model—to think about the description of resources. It was developped in 1999 and is a W3C standard. Its original name was "metadata data model".

The main idea behind RDF is to make statement about resources. RDF defines a statement to be anything which has a *Subject*, a *Predicate* and an *Object*. Due to this, a statement is also refered to as a *triple*. Let's first examine each of its constituents.

Subject The subject is the *resource* to be described by the statement.

Predicate is the trait or the aspect of the resource we want to describe.

Object is the value attributed to the subject through the predicate. It can either be a *resource* or a *literal*, which is a textual value.

3.1.1 Representation

Since RDF statements are triples, it is possible to represent them graphically using a *labeled, directed multi-graph*. Or, in laymen terms, using arrows, boxes and ellipses.

Let's have an example. Let's see how we could make RDF statements about Shrodinger, the turtle that I don't really have.

Figure 3.1 is an RDF graph. It is *labeled* and *directed*. There are ellipses, arrows and boxes. This graph makes statements about a resource—*http://sajrmajaz.zlzpp.com/shrodinger*. Look at the arrows. They each represent a *predicate*, an attribute if you will. The node on the pointy side of the arrow is the *object* of the statement. In the case of the *rdf:type* predicate, the

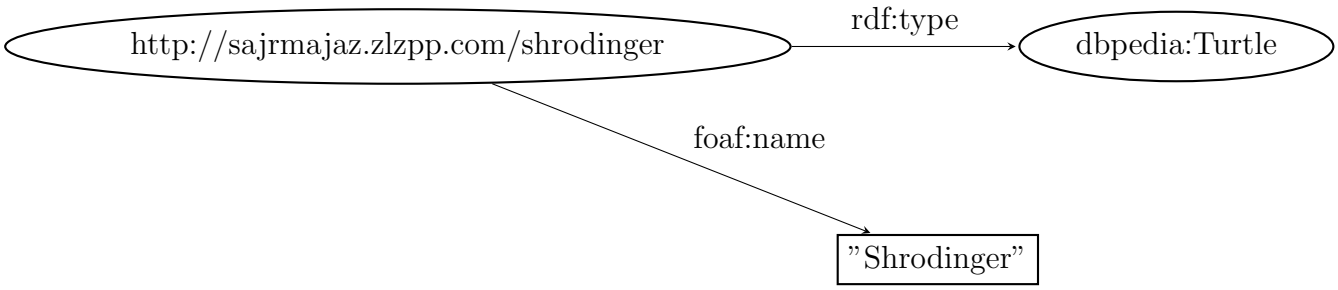

Figure 3.1: Some RDF statements

object is a resource—it is represented by an ellipse. In the case of *foaf:name*, it is a *literal*.

The *rdf:type* and *foaf:name* labels are shortcuts. Since they are resources, they are represented by a URI. However, the graph wouldn't be readable if we wrote the full URI of each resource. So the convention is to declare namespaces, such as *rdf:* and *foaf:*. They simply expand to the full URI of the vocabulary they represent. I could have declared my own namespace, *zlzpp:*, to represent *http://sajrmajaz.zlzpp.com/*, so that I could write *zlzpp:shrodinger* to refer to the subject.

3.1.2 Textual syntax

I won't go over the benefits of using a visual representation for RDF. It works really well, but somehow falls short when there is an important amount of statements. Not to mention that computer are really good at processing text. For that, very intelligent people sat together and worked on ways to express RDF using text only: on *serialization*.

3.2 A SPARQL of hope

Data represented as RDF statements sit there in graphs, ready to be called upon. One way of using these data is to dereference the URI of the resource we are interested in. But most of the time, we want to get *all the resources of type foaf:Person*, or all the *Bands who play Jazz music*. Instead of wanting data about resources, we want to find the resources themselves.

Since data live in a *data silo* (or graph), these silo sometime provide a

special service calld an *end-point*. Like the yellow pages, end-points give you information about resources present in the silo.

In order to use that service—if such a service exists in the data silo you wish to access—, you need to use a language to communicate with it, the *Sparql Protocol and Query Language*, *SPARQL*.

3.3 Meeting resources

There we are, finally ready to meet our first resources. How exhilarating!