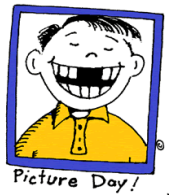
Web Services and Web Data

XJCO3011

Session 14 – Linked Data

A Motivating Example

How can you answer the question: Did my teacher go to Finland? by examining his profile page on the university website?



Personal Page

University Staff Profile Page

Hobbies:
Skiing

Facebook

Personal Web Page

During My Recent Trip to Ruka

Facebook Page

Google

Ruka

Google Search

Ski Resort | Ski.ruka.fi

Google Search Results

---------- Finland ------

Ruka Ski Resort

The World Wide Web

- A collection of linked documents that are full of data.
- Many of these documents have little, if any, structure imposed on the data (mostly images and free text).
- The data is available in so many formats such as HTML, XML, PDF, TIFF, CSV, Excel spreadsheets, embedded tables in Word documents, and many forms of plain text.
- This kind of data has a limitation: it's formatted for human consumption.
- It often requires a specialized utility to read it.
- It's not easy for automated processes to access, search, or reuse this data.
- Further processing by people is generally required for this data to be incorporated into new projects or allow someone to base decisions on it.
- With the exception of some very simple cases, only humans can analyse the semantic relationships between data in various web pages.

The Linked Data Web

- Linked Data refers to a set of techniques for publishing and connecting structured data on the Web
- It adheres to standards set by the World Wide Web Consortium (W3C).`

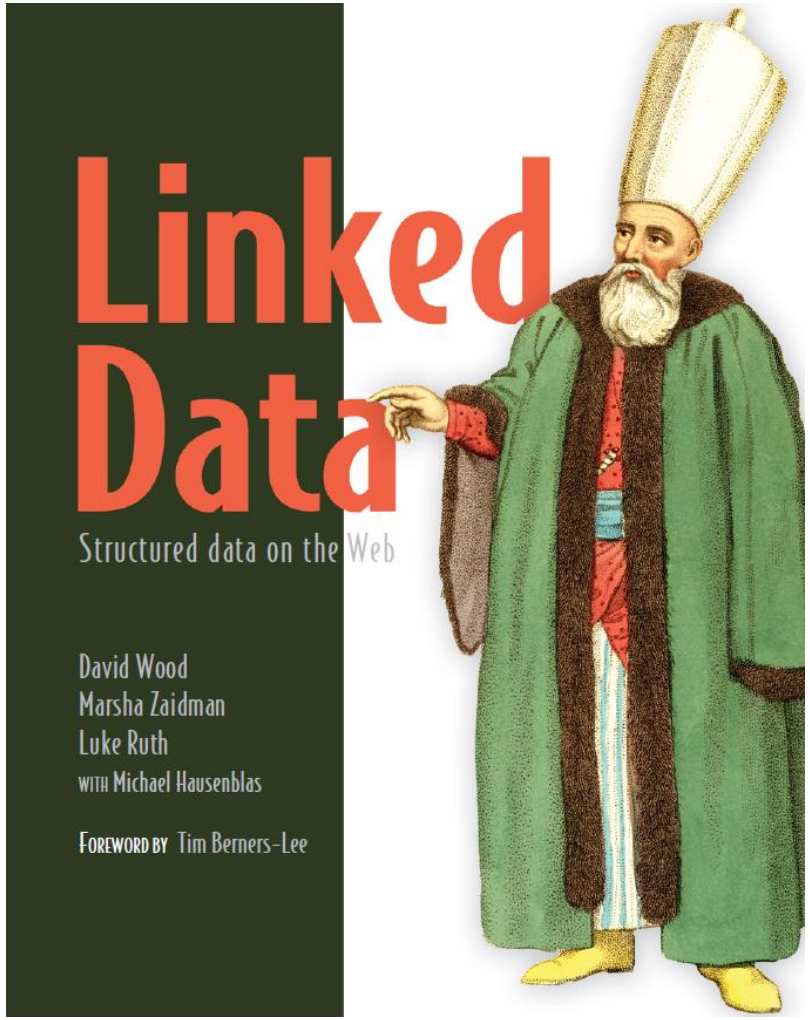
One of the persons is "Anakin," known as "Darth Vader" and "Anakin Skywalker." He has a wife named Padme Amidala.

Unstructured Data. Good for Humans

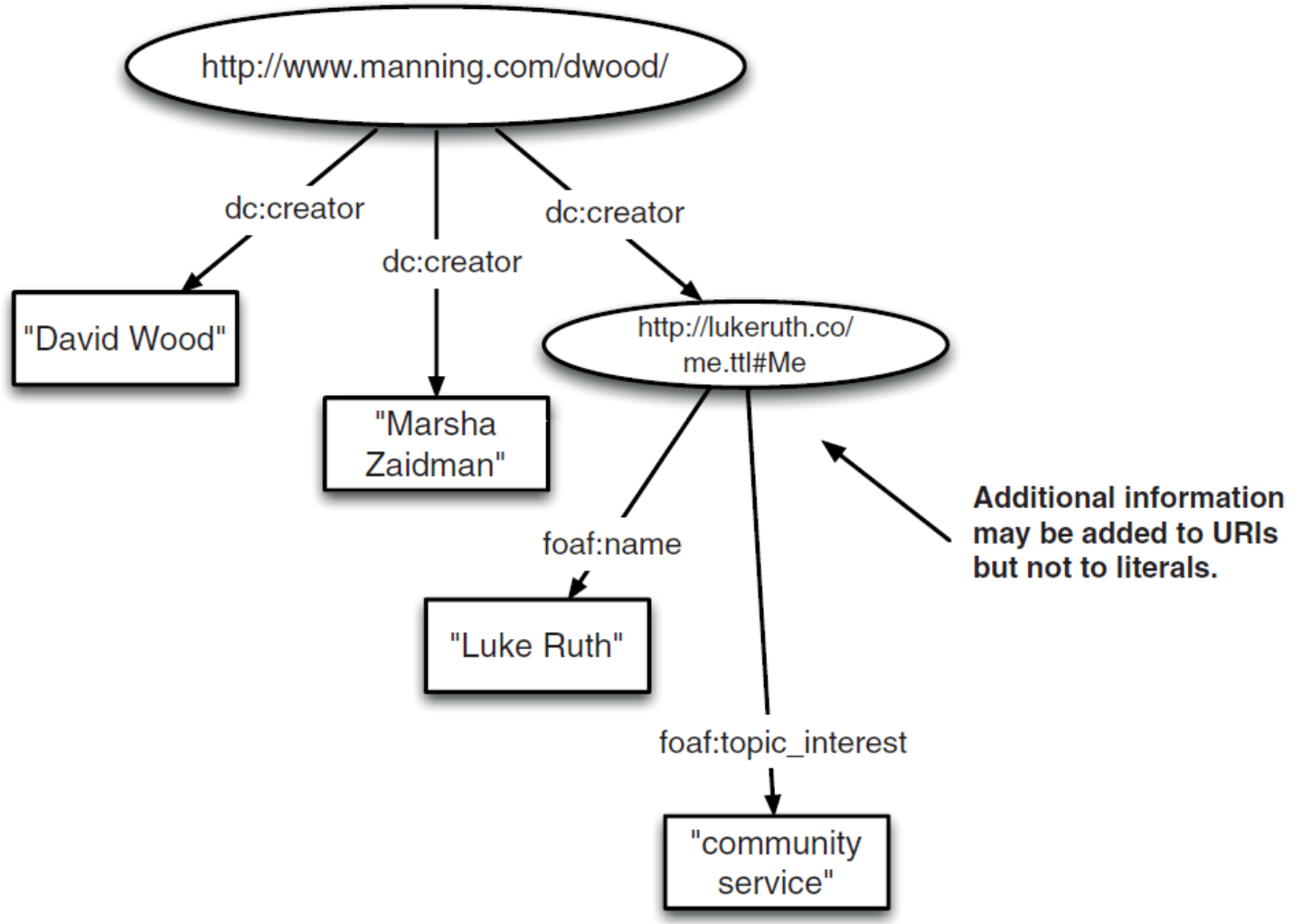
```
@base <http://rosemary.umw.edu/~marsha/starwars/foaf.ttl#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rel: <http://purl.org/vocab/relationship>.
@prefix stars: <http://www.starwars.com/explore/encyclopedia/characters/> .
 <me> a foaf:Person;
        foaf:family_name "Skywalker";
        foaf:givenname "Anakin";
        foaf:nick "Darth Vader";
        rel:Spouse_Of <stars:padmeamidala/> .
```

Structured data (RDF Turtle format). Good for Automated Agents

The Linked Data Web



Unstructured Data (Picture)



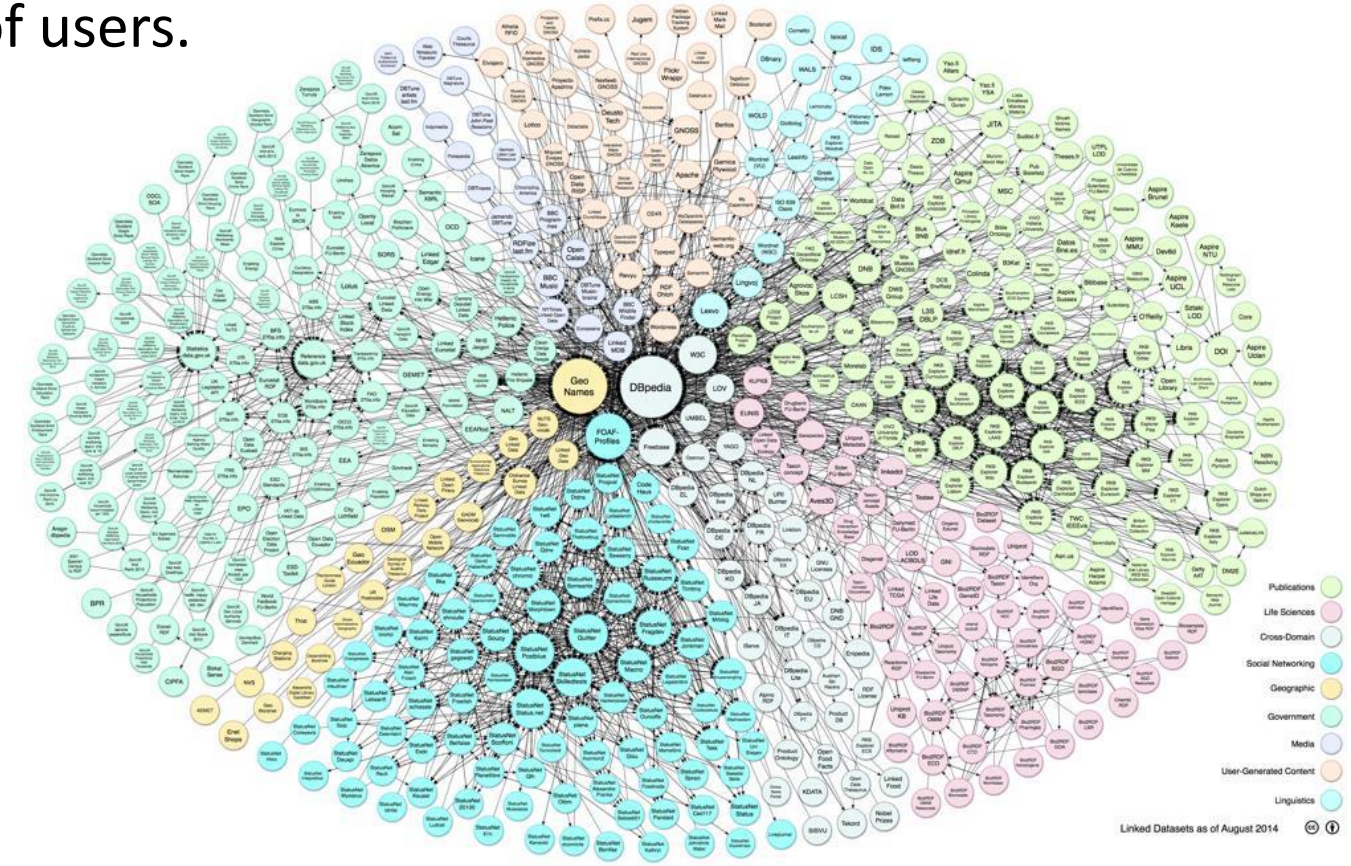
Linked Data (Structured)

DBpedia

- The DBpedia project (http://dbpedia.org) extracts factual information from Wikipedia articles and publish them, on the Web, as structured data.
- Fortunately, most of the content in Wikipedia is highly structured (e.g. the "infobox" in the upper right of a Wikipedia page).
- The 2016-04 release of the DBpedia data set describes 6 million entities, including 1.5 million persons, 800,000 places, 135,000 music albums, 100,000 films, 300,000 species and 5,000 diseases.
- This dataset is open and may be explored by anyone to extract or create new knowledge.

The Semantic Web

- When the elements of the web are structured data entities, connected to each other with semantic relations, the resulting graph is called the semantic web.
- This enables automated agents to access the Web more intelligently and perform more complicated queries on behalf of users.
- WWW: emphasizes the connection between people and information
- The semantic web: helps machines understand the relationships and meanings between pieces of information



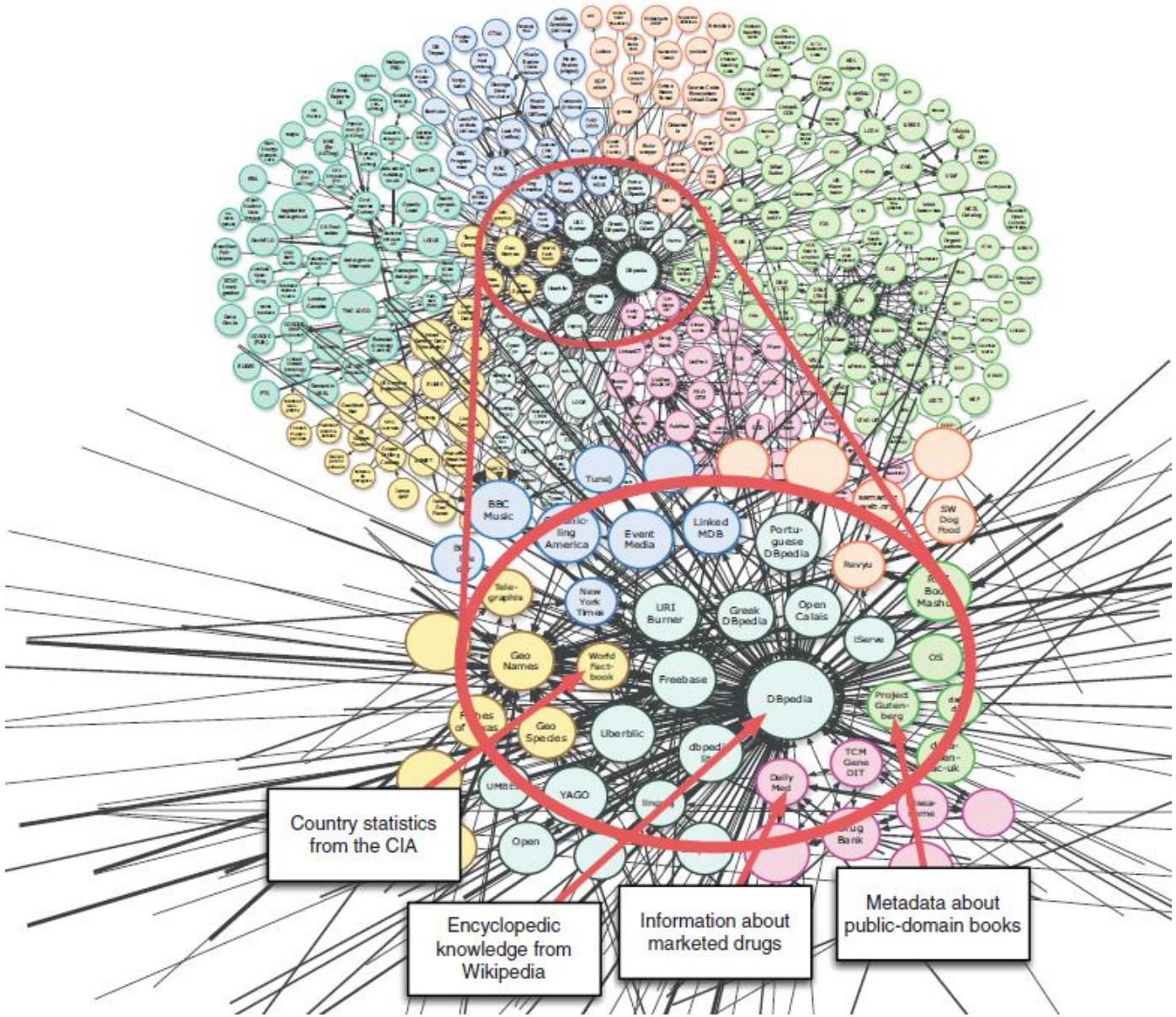
The Linking Open Data project

- There exist a huge body of datasets that are open and freely available on the Web. These open-content projects are as diverse as encyclopedias, dictionaries, government statistics, chemical and biological collections, endangered species, bibliographic data, music, artists, songs, academic research papers. They are all available through the same data format (RDF). This is all due to the Linking Open Data (LOD) project.

- The Linked Open Data (LOD) project is a community activity started in 2007 by the W3C's Semantic Web Education and Outreach (SWEO) Interest Group. The project's goal is to "make data freely available to everyone."

- This collection of Linked Data published on the Web is referred to as the LOD cloud. An attempt to visualizing the LOD cloud is shown in the next slide.

The Linking Open Data project



- The Linked Open Data cloud in late 2011.
- The circles represent freely available datasets and the arrows represent links between them.
- Some quick facts regarding the LOD cloud:
  - The LOD cloud has doubled in size every 10 months since 2007 and currently consists of more than 300 datasets from various domains. All of this data is available for use by developers!
  - As of late 2011, the LOD cloud contained over 295 datasets from various domains, including geography, media, government, and life sciences. In total the LOD cloud contained over 31 billion data items and some 500 million links between them.
  - etc.…

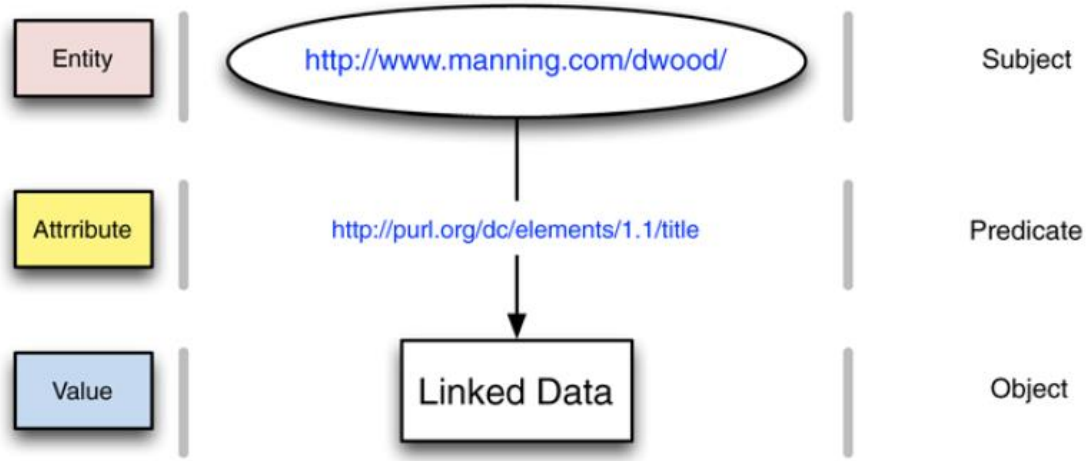
Quick Facts About the LOD Cloud.

- The LOD cloud has grown so large that no attempt was made to visualize after 2011.
- More than 40% of the Linked Data in the LOD cloud is contributed by governments (mainly from the United Kingdom and the United States), followed by geographical data (22%) and data from the life sciences domain (almost 10%).
- Life sciences (including some large pharmaceutical companies) contribute over 50% of the links between datasets.
- Publication data (from books, journals, and the like) comes in second with 19%, and the media domain (the BBC, the New York Times, and others) provides another 12%.
- The original data owners themselves publish one-third of the data contained in the LOD cloud, whereas third parties publish 67%. For example, many universities republish data from their respective governments in Linked Data formats, often cleaning and enhancing data descriptions in the process.

The Resource Description Framework (RDF)

- Linked Data uses RDF as a data model.
- A single RDF statement describes *two things and a relationship between them*.
- This is called an Entity-Attribute-Value (EAV) data model
- Linked Data people often call the three elements in a statement the subject, the predicate, and the object.

	A	B
1	id	title
2	http://www.manning.com/dwood/	Linked Data
3		



The Elements of RDF Statements

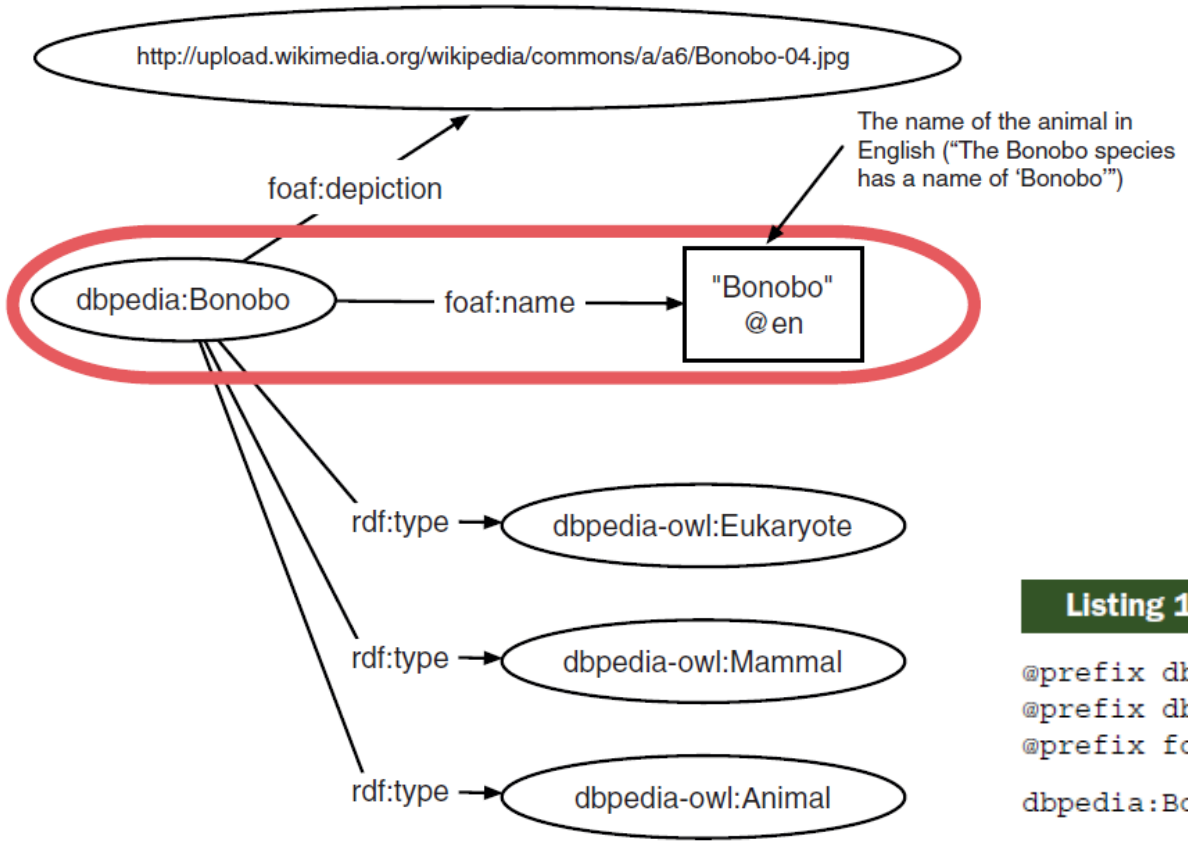
- An entity (or subject) is anything that we can name by a URI, such as a person, a book, a car, or a web page. In the previous case, the subject is the URI that uniquely identifies the book.
- An entity's attribute (or predicate) relates the subject to another entity or provides information about the entity itself (which we call a property of the subject). In the previous case, we're using a relationship for the book's title as the predicate and the book's title as the object.
- Through this standardized data model, an API is created that's consistent over all Linked Data sources. You only have to learn the Linked Data model once and you can use any kind of data source that complies with it.
- Anytime you want to know what a predicate means, you can type its URI into a web browser and look for information about it.

RDF Example

<http://example.com/my_temperature_data> rdfs:label "Temperature observations";
rdfs:comment "Temperature observations at Galway Airport";

- <http://example.com/my_temperature_data> is a URI representing a sample spreadsheet of temperature data, which forms the entity (subject) of an RDF statement.
- The two components in rdfs:label "Temperature observations"; are the attribute (predicate) and the property (object) of the first RDF statement. In this case, we're saying that the spreadsheet may be given a human-readable label of "Temperature observations".
- rdfs:comment "Temperature observations at Galway Airport"; provides another attribute and property for the same subject, which forms another RDF statement.
- We can keep adding information about our spreadsheet that way until we're finished.
- There's no restriction in RDF about what you can link to or describe. RDF statements create graphs of metadata. We often use the term RDF graph because of this.

RDF Diagrams



RDF data snippet

A Bonobo

Listing 1.2 Excerpt of the Linked Data about bonobos in Turtle format

```
@prefix dbpedia:        <http://dbpedia.org/resource/> .
@prefix dbpedia-owl:    <http://dbpedia.org/ontology/> .
@prefix foaf:           <http://xmlns.com/foaf/0.1/> .

dbpedia:Bonobo    rdf:type    dbpedia-owl:Eukaryote ,
                  dbpedia-owl:Mammal ,
                  dbpedia-owl:Animal .

dbpedia:Bonobo    foaf:name    "Bonobo"@en ;
                  foaf:depiction <http://upload.wikimedia.org/wikipedia/
     commons/a/a6/Bonobo-04.jpg> ;
```

Name of the animal in English ("A Bonobo has a name of 'Bonobo'")

- A useful link defines a textual syntax for RDF called Turtle that allows an RDF graph to be completely written in a compact and natural text form : https://www.w3.org/TR/turtle/#grammar-production-predicateObjectList

That's it Folks

Further Reading

Chapter 1 and 2: Linked Data Structured Data on the Web, David Wood et al.