Future Internet

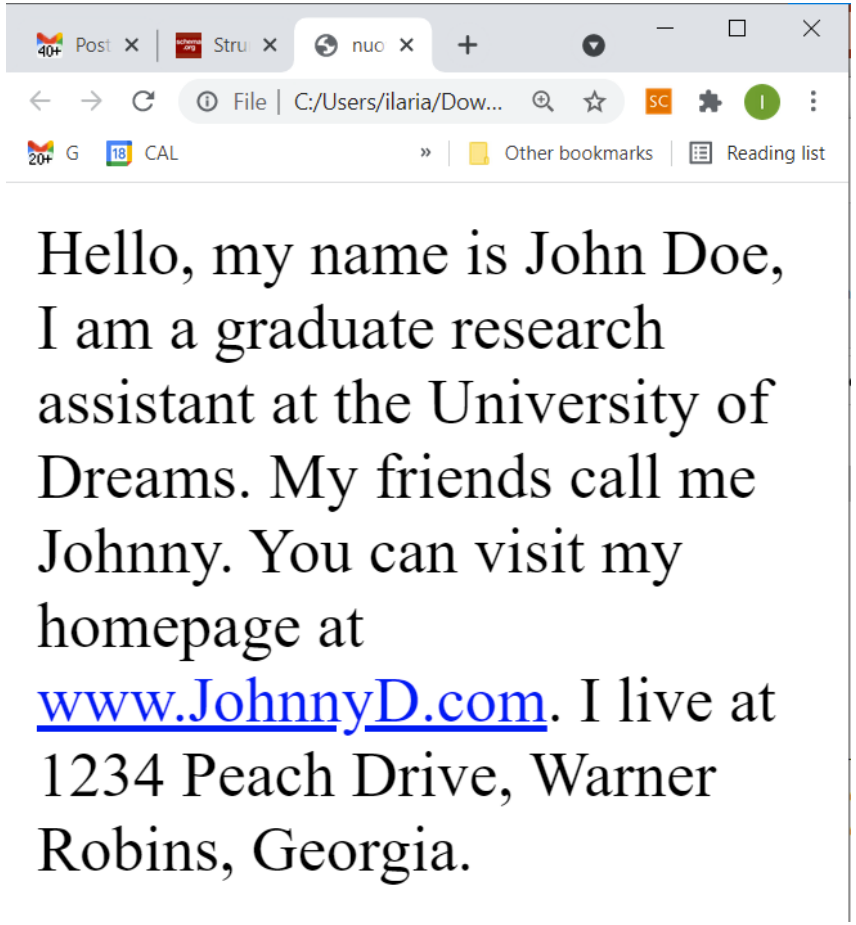
ILARIA TORRE

Ilaria.torre@unige.it

Semantic annotation of web pages

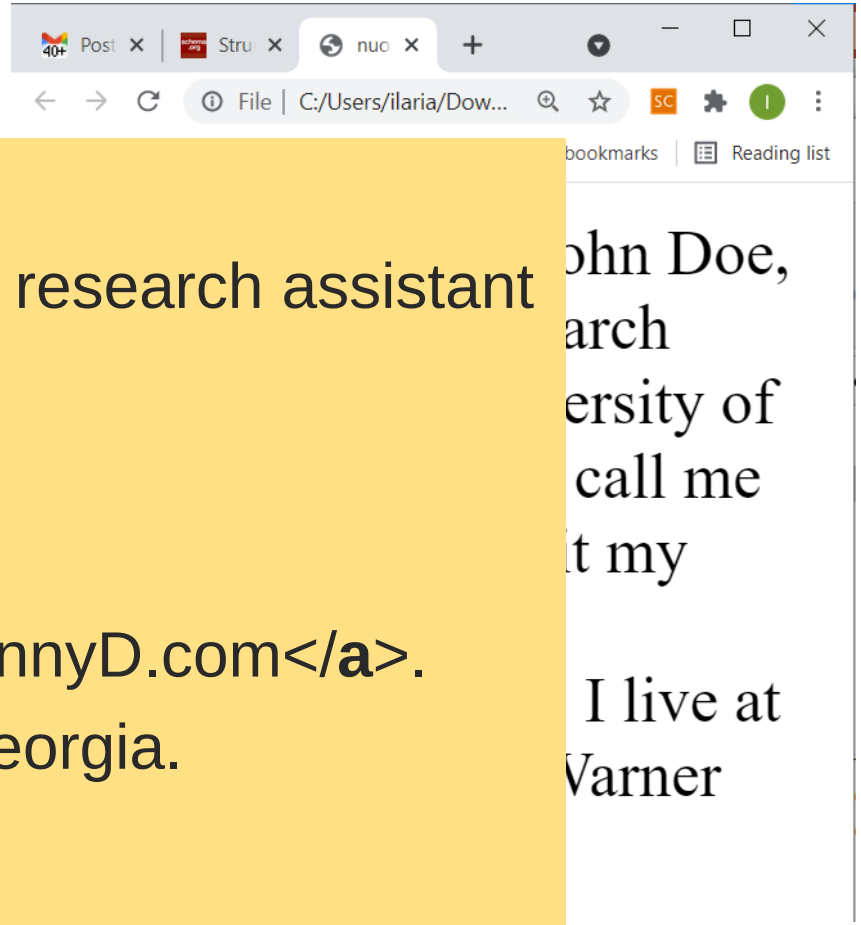
An example: how to annotate this web page?

Goal: enabling content to be unambiguous and **machine-understandable**



Source: https://www.w3.org/wiki/Microdata_RDFa_Merge

HTML code




```
<div>
Hello, my name is John Doe, I am a graduate research assistant
at the University of Dreams.
My friends call me Johnny.
You can visit my homepage at
<a href="http://www.JohnnyD.com">www.JohnnyD.com</a>.
I live at 1234 Peach Drive, Warner Robins, Georgia.
</div>
```

Source: https://www.w3.org/wiki/Microdata_RDFa_Merge

Steps to annotate web pages using RDF-based structured data

Steps:

Identify the goals of semantic annotation: What are the software agents that will use this data? For example:

- Is it useful for search engines?
- For well-known mashup and integration services?
- For possible services not known?

  In the following example, we want to make all text content machine-processable (so we'll skip this point)

Based on your goals, determine what to annotate by pre-processing the content

Identify the vocabulary/ontology for the annotation (if not available, extend an existing one or create a new one by defining the alignment criteria with the existing ones)

Choose the annotation language based on the objectives and expressive power required

1° Identify what to annotate, express statements as triples, make explicit what is implicit

The web page text is a sequence of statements. They can be represented as a sequence of RDF triples (S,P,O).
But, implicit knowledge has to be made explicit

- [I am a **Person**] (1° statement)
- my name is John Doe (2° statement)
- I am a graduate research assistant (3° statement)
- [I am] at the University of Dreams (4° statement)
- my homepage www.JohnnyD.com (5° statement)
- I live at [address] (6° statement)
- [This is an **Address**] (7° statement)
- [street is] 1234 Peach Drive (8° statement)
- [locality is] Warner Robins (9° statement)
- [region is] Georgia Robins (10° statement)

For a user this information is implicitly known; a software agent needs it explicitly

The same as above

2° Group the assertions that concern each subject

The main logical subject:

[the subject is a **Person**] (1° statement)

name John Doe (2° statement)

graduate research assistant (3° statement)

University of Dreams (4° statement)

homepage www.JohnnyD.com (5° statement)

live at [address] (6° statement)

The logical subject of the embedded clause:

[the subject is an **Address**] (7° statement)

[street] 1234 Peach Drive (8° statement)

[locality] Warner Robins (9° statement)

[region] Georgia Robins (10° statement)

3° Chose a vocabulary/ontology to annotate content

In our examples we will use schema.org

For the 1logical subject we will use type **Person** and its **properties**

*<1logical subject, typeof, **Person**>*

*<1logical subject, **name**, John Doe>*

*<1logical subject, **additionalName**, Johnny>*

*<1logical subject, **url**, http://www.johnnyd.com/>*

etc.

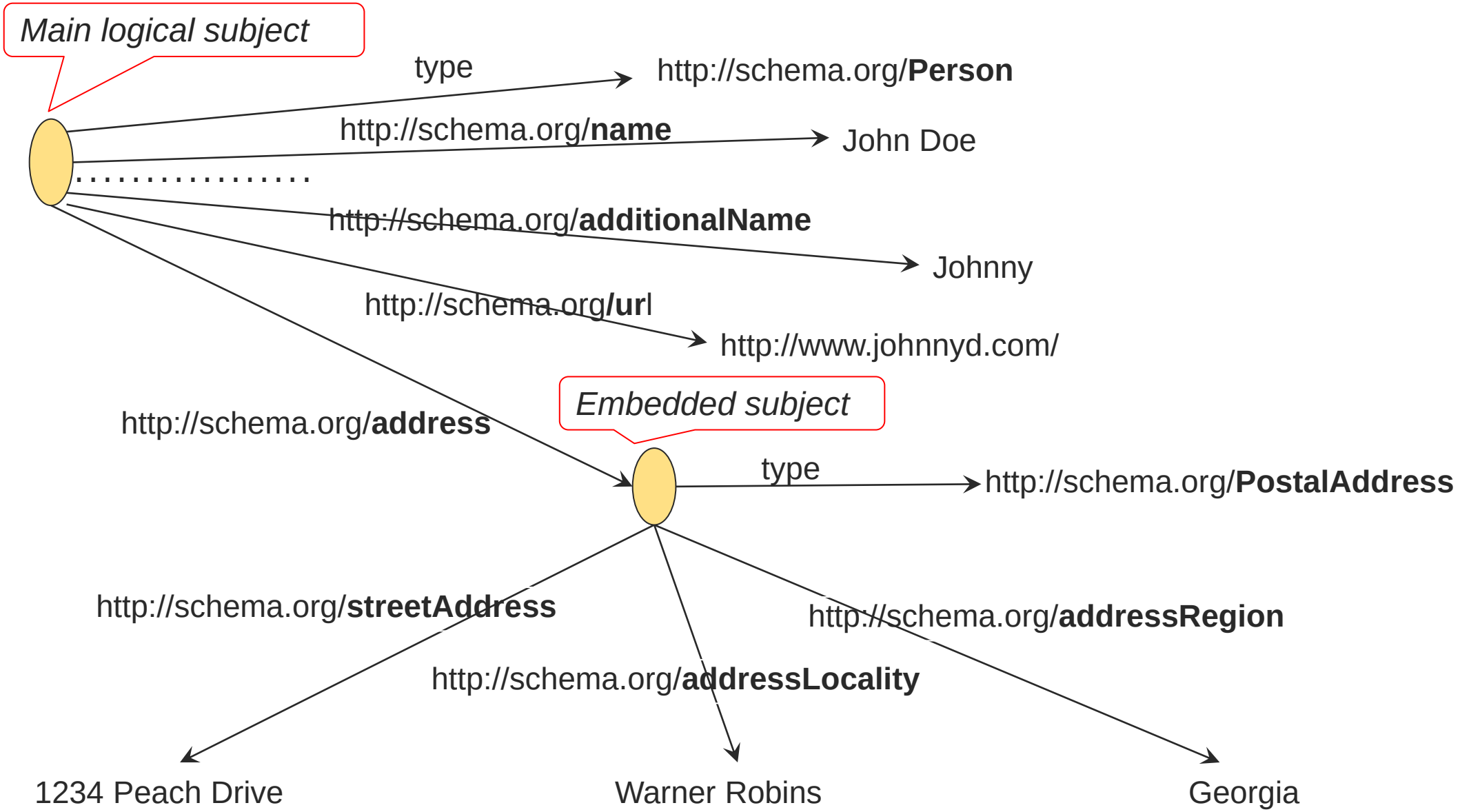
4° Annotate content

Subject → Predicate → Object

- RDF graph
- Serialization

4- RDF graph

Subject → Predicate → Object



4- serialization

Representing the RDF graph using one of the serialization formats.

We will use JSON-LD

Lab: try and practice

- Objective 1: Verify "practically" that [almost] any sentence can be expressed via RDF triples
  - Choose a portion of any web page to annotate of your choice
  - Copy the displayed text in natural language and turn it into triples by following steps 1 and 2

- Objective 2: Learn how to use a vocabulary/ontology to semantically annotate data
  - Go to schema.org
  - Among the **Type** (classes) identify the ones that can best represent the subjects of the triples that you identified in the previous step
  - Click on each Type and scroll through the properties to find the ones that best express the properties of the subject expressed in the natural language sentence

Lab: try and practice

- Note: the **Expected type** is the object in the triple RDF.
- Be careful distinguishing between data properties and object properties, based on the Expected type

➢ Objective 3: Represent triples as RDF Graph

- On paper or using diagram software, draw the RDF graph using the Types chosen by schema.org as the classes to which the Subjects belong and using the properties of those subjects to define the edges of the graph, following the model in Step 4