Knowledge Graphs

Oscar Romero

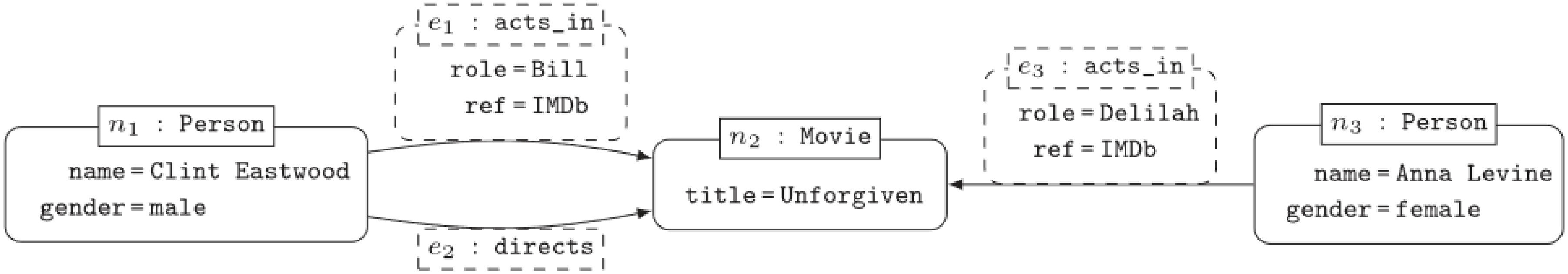
Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya

New Challenges for Data Modeling

- Flexible and generic means to represent data
  - Semi-structured data models
- Data exchange and exploitation
  - Semantics (metadata) as first-class citizen
  - Linked data (i.e., with semantic *pointers*)

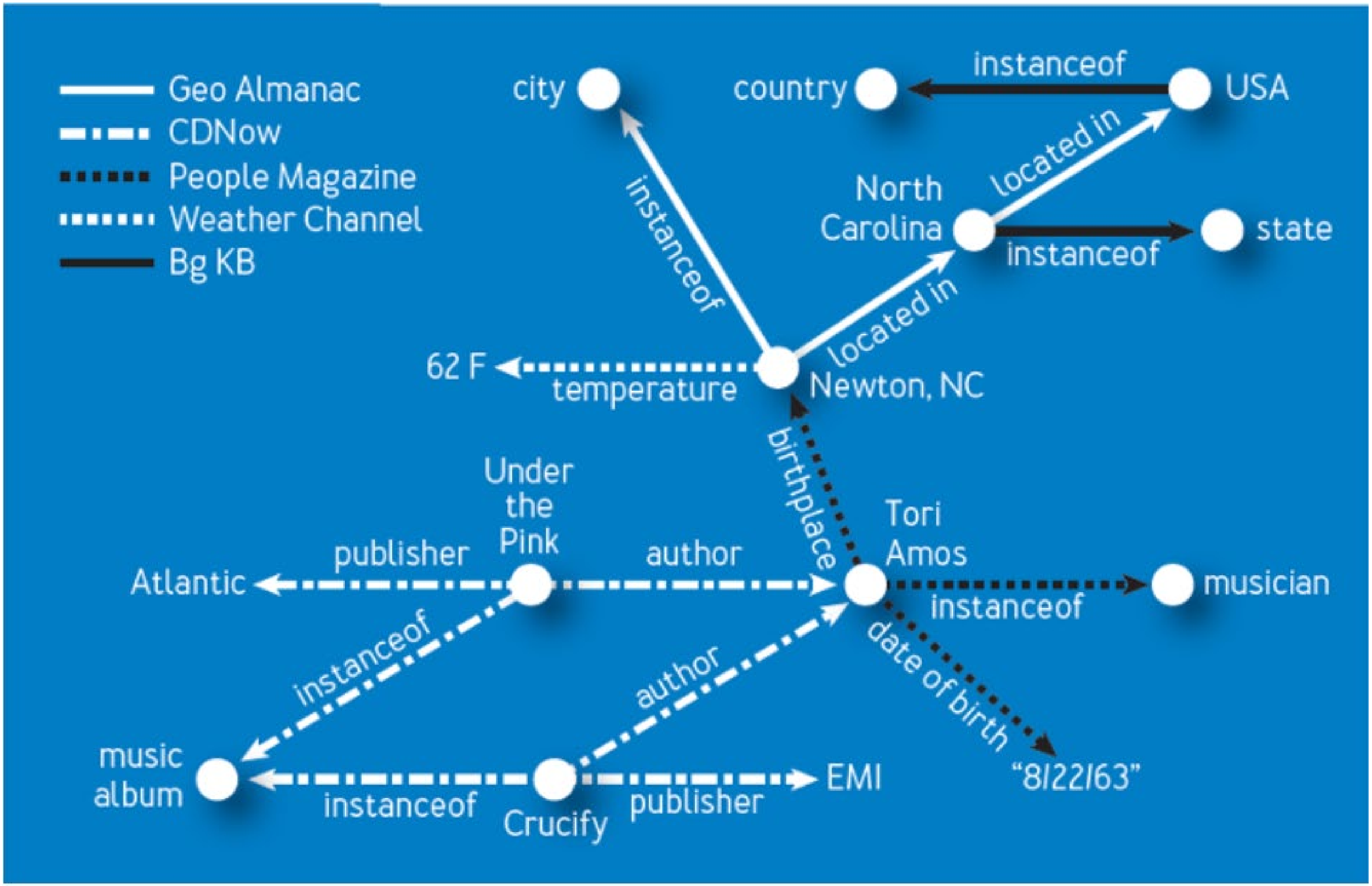
An example of Property Graph:



Sharing Data

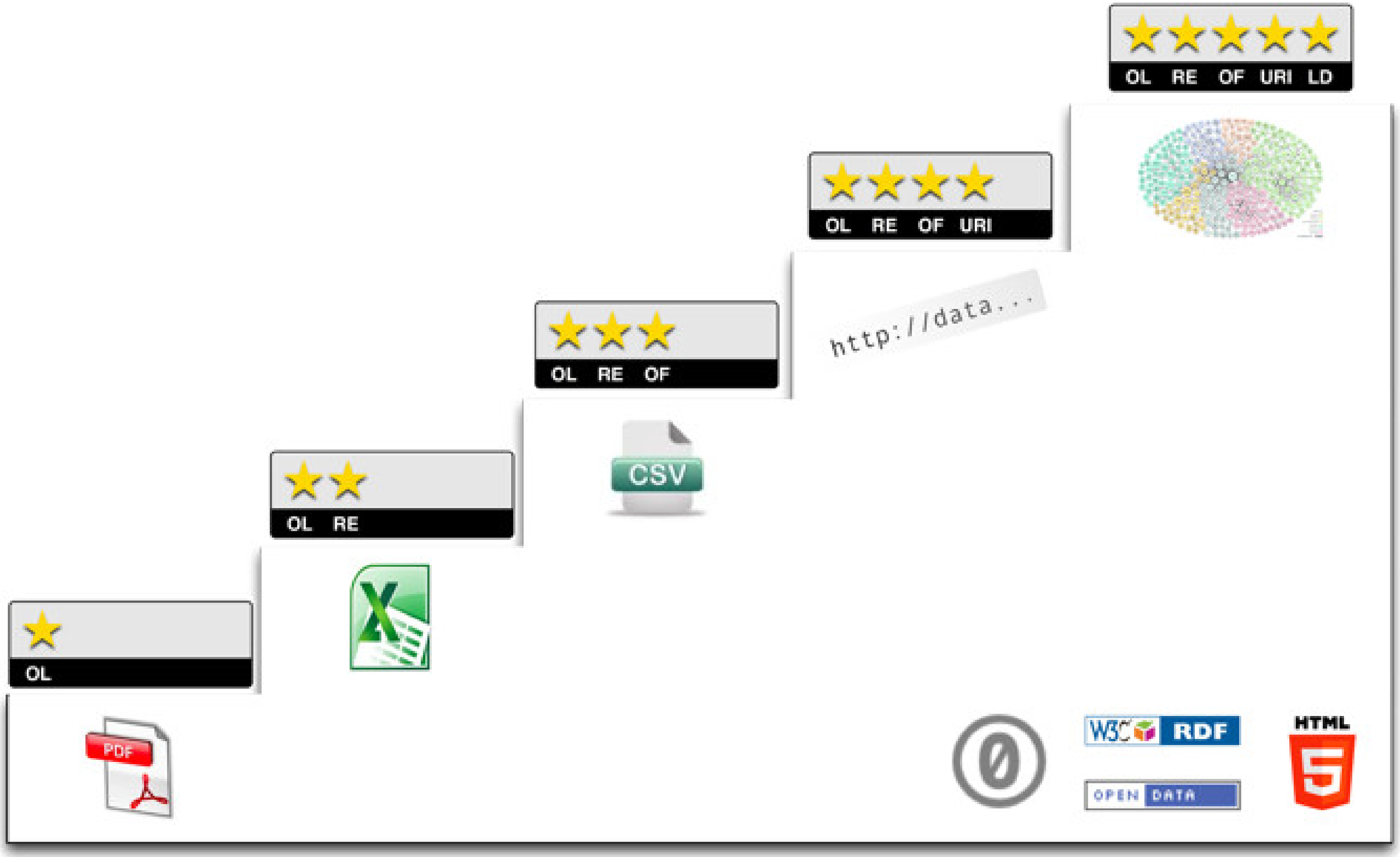
- New Challenges
  - Flexible and generic means to represent data
    - Semi-structured data models
  - Data exchange and exploitation
    - Semantics (metadata) as first-class citizen
    - Linked data (i.e., with semantic *pointers*)
- Property graphs do not provide means to define semantic pointers
- The Knowledge Representation community made the biggest steps in this direction
  - Most powerful semantic models are graph-based
  - A semantic repository is a **distributed** semantic database

The Envisioned Idea



Extracted from: https://queue.acm.org/detail.cfm?id=2857276

Is Your (Open) Data 5 Stars?



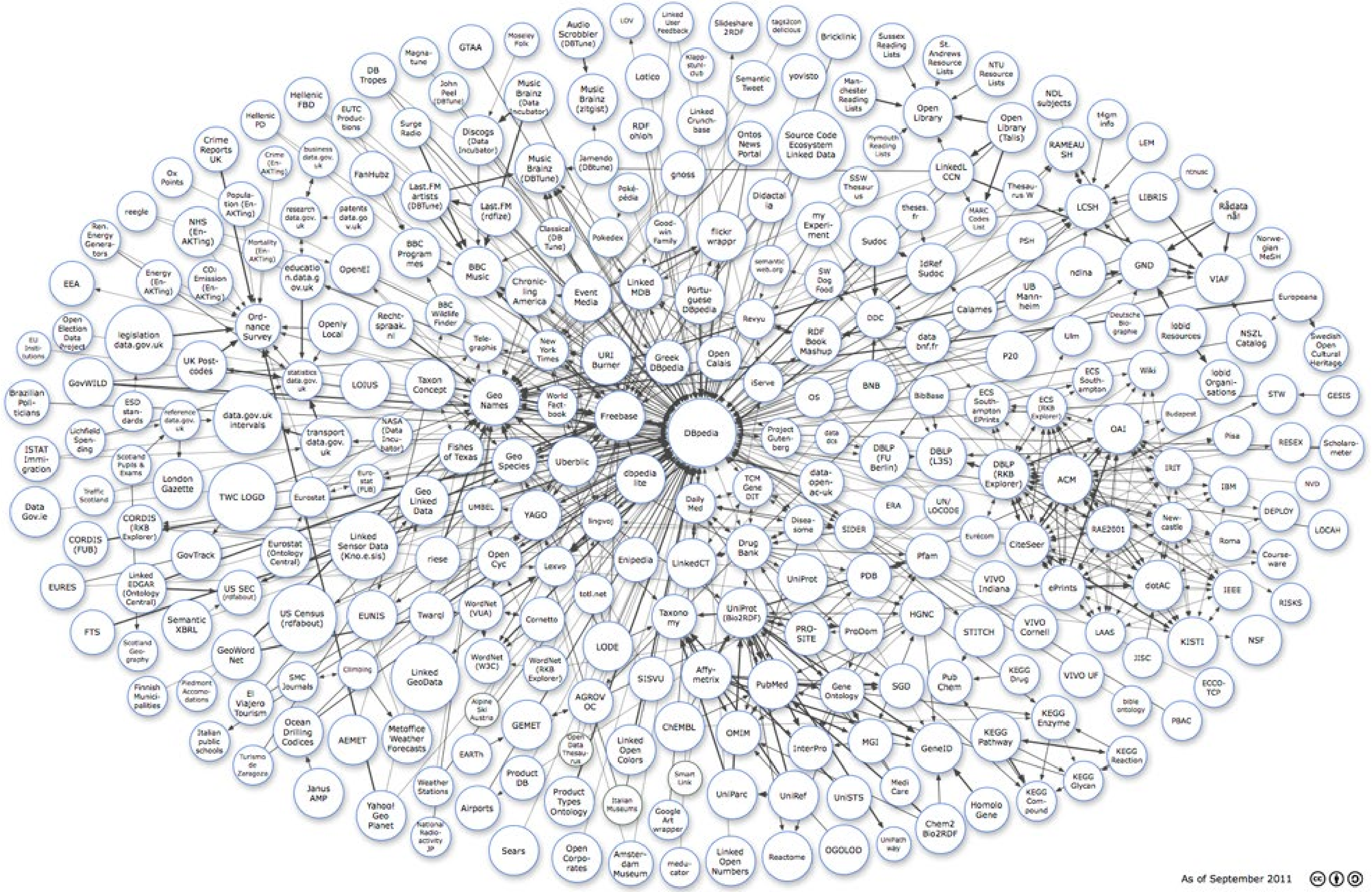
Semantic Data Models

- Conceptual Data Models
  - Ontologies **LINKED DATA**
- Logical Data Models
  - XML **DATA SILOS**
  - JSON **DATA SILOS**
  - RDF **LINKED DATA**

Linked Data (How To)

- As it was described, Linked Data sits at the conceptual level
  - Lack of standards!
- Linked Data So Far (At the logical level)
  - W3C Standards for *sharing meaning*
    - RDF (Resource Description Framework)
      - Subject predicate object
    - OWL / Ontologies
      - Common conceptualizations
  - URIs (*global identifiers*)
    - URN (Universal Resource Name) ~ id
    - URL (Universal Resource Location) ~ *where* to locate it
      - Dereference (HTTP protocol)

The Linking Open Data Project



As of September 2011

"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

Schema.org: Example

- Schema.org is a global initiative to mark up (i.e., attach semantics to) data
  - It provides a vocabulary of terms (concepts) and their relationships: https://schema.org/docs/full.html
- Example:

```
<div itemscope
itemtype="https://schema.org/Offer">
<span itemprop="name">Blend-O-Matic</span>
<span itemprop="price">$19.95</span>
<link itemprop="availability"
href="https://schema.org/InStock"/>Available
today! </div>
```

The element markup starts by typing it

State properties (attrs) of the element

Data values can be constrained with enumerations

- Google (and others) have built their semantic-aware searchers based on schema.org

Knowledge Graphs

RDF

Disclaimer

- Even if knowledge graphs were born in the Semantic Web (c.f., to fulfill the concept of Linked Data) we are going to use them for a broader range of purposes
- **DO NOT UNDERSTAND RDF AS SYNONYM OF THE SEMANTIC WEB**

RDF

- RDF: Resource description format
    - Resources or objects (identified by a URI),
    - Literals (atomic values such as string, dates, numbers, etc.) and
    - Properties (binary relationships between resources and literals)
- The basic RDF block is the ***<u>triple</u>****:* a binary relationship between two resources or between a resource and a literal
    - *<Subject predicate object>*
        - *Subject* S has value *object* O for *predicate* P
        - Subject and predicate must be URIs
        - Object can be a URI or a literal (i.e., a constant value)
    - It accepts blank nodes (whose URL is _:)
    - The resulting metadata can be seen as a graph
- It is a simple language for describing *annotations* (facts) about resources identified by URIs
    - The most basic **ontology** language
        - Triples map to FOL as grounded atomic formulas (subject and object are constants)
        - Blank nodes map to existential variables
- SPARQL is the de facto language to query RDF and its variants

RDF: The Big Picture

- Semantics

http://www.w3.org/TR/rdf11-primer/

  - **Triples** (i.e., RDF statements),
  - To create **semantic graphs** (aka RDF graphs or knowledge graphs)

- Syntax
  - **XML-based serialization** (aka RDF syntax)
    - The way to express RDF triples in a machine-processable format

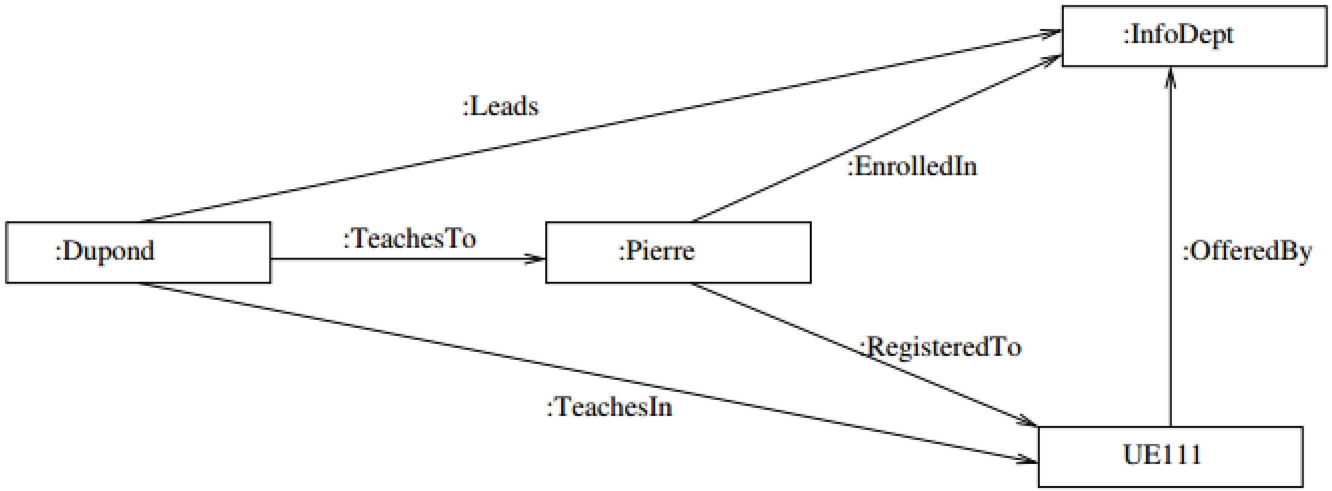
RDF Example (I)

- RDF triplets or statements:

⟨ :Dupond :Leads :CSDept ⟩
⟨ :Dupond :TeachesIn :UE111 ⟩
⟨ :Dupond :TeachesTo :Pierre ⟩
⟨ :Pierre :EnrolledIn :CSDept ⟩
⟨ :Pierre :RegisteredTo :UE111 ⟩
⟨ :UE111 :OfferedBy :CSDept ⟩

Subject	Predicate	Object
:Dupond	:Leads	:CSDept
:Dupond	:TeachesIn	:UE111
:Dupond	:TeachesTo	:Pierre
:Pierre	:EnrolledIn	:CSDept
:Pierre	:RegisteredTo	:UE111
:UE111	:OfferedBy	:CSDept

- RDF graph:



Extracted from Web Data Management (Abiteboul et al.)

RDF Example (II)

- RDF is typically serialized as XML (syntax)
- The rdf URL is a namespace for RDF
  - The URI is: http://www.w3.org/1999/02/22-rdf-syntax-ns#

- Example:

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cd="http://www.recshop.fake/cd#">

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Empire Burlesque">
  <cd:artist>Bob Dylan</cd:artist>
  <cd:country>USA</cd:country>
  <cd:company>Columbia</cd:company>
  <cd:price>10.90</cd:price>
  <cd:year>1985</cd:year>
</rdf:Description>
```

- Other RDF Syntaxes:
  - **Turtle** (human-readable)
  - N-triples
  - Notation 3
  - …

Turtle:

```
(…)
<#empire-burlesque>
      cd:artist    <#Bob-Dylan>   ;
      cd:country <#USA>           ;
      cd:price     10.90          .
(…)
```

https://www.w3.org/TR/turtle/

Extracted from W3C School

Creating (Sharing) RDF Datasets

- RDF Datasets follow these principles:
  - Every single piece of data created…
    - Is machine-readable,
    - Its meaning is explicitly defined,
    - Its linked to other external data sets,
    - And can be linked to from other external data sets
- How? Tim Berners-Lee outlined a set of rules:
  - Use URIs as names for things
    - Universal identifiers to represent real-world objects
  - Use HTTP URIs so that people can look up those names
    - Universally available (where to locate it)
  - When someone looks up a URI, provide useful information, using RDF and SPARQL (DESCRIBE clause)
    - It describes the object by listings its features and relationships
  - Include links to other URIs, so they can discover more things
    - Relationships as first-class citizens (information integration)
- Consequences:
  - Data is separated from formatting
  - Data is self-describing
    - Dereferencing to deal with unknown vocabularies (i.e., metadata from existing RDF datasets)
  - HTTP as standard to locate entities

Most Popular RDF Vocabularies

- Some organizations, groups, etc. created some RDF vocabularies (i.e., RDF graphs), which are available on-line (RDF namespaces). Some examples:
  - Simple knowledge organization system (SKOS) http://www.w3.org/TR/skos-reference/
  - Friend-of-a-friend (FOAF) http://xmlns.com/foaf/spec/
  - Statistical Data and Metadata Exchange (SDMX) http://sdmx.org/
  - The RDF Data Cube Vocabulary (QB) http://www.w3.org/TR/vocab-data-cube/
  - Geospatial resources (GeoNames) http://www.geonames.org/ontology/documentation.html

Activity: Modeling in RDF

- *Objective: Grasp the idea behing RDF modeling*
- (20’) Model a RDF Graph capturing data about lecturers, courses and students:
  - *A student enrolls several courses from his faculty per semester. He is forced to enroll, at least, one course per semester. Each course has one responsible lecturer but potentially, it might have several lecturers.*
  - Your graph must be a correct RDF graph and therefore defining the namespaces used
- Afterwards, write the triples you created in Turtle syntax

Basics on RDF Modeling

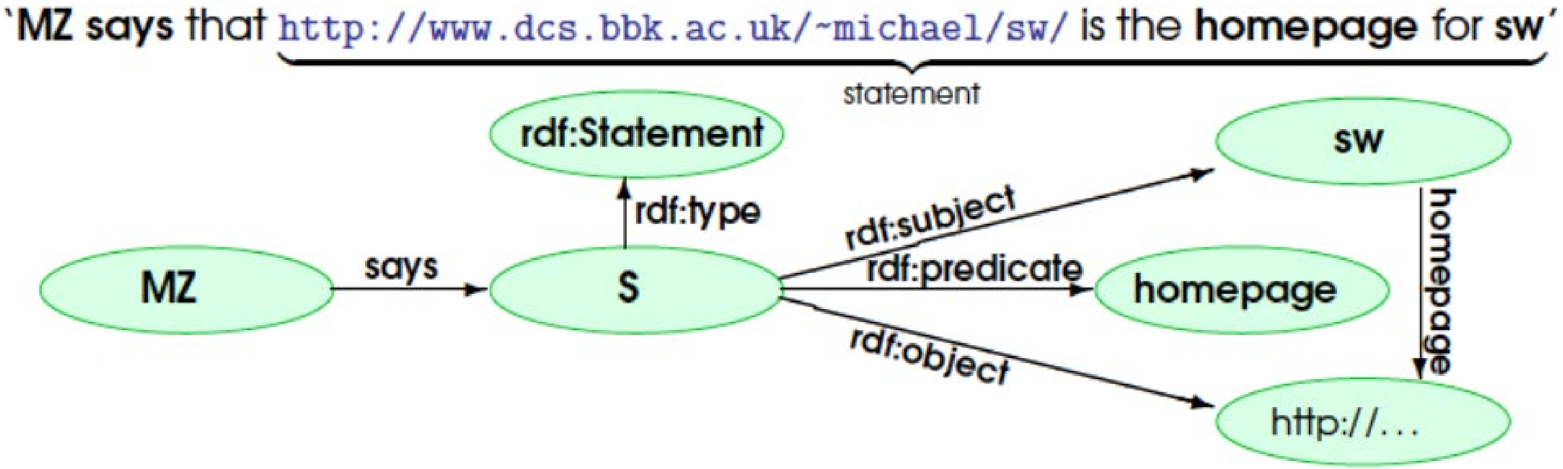
- RDF modeling is based on binary relationships
  - Since n-ary relationships may be needed, blank nodes were presented as solution
- We cannot express neither schema (not even the concept of label) nor additional constraints
  - Example: *at least one, at most three, the domain of a property must be of type X*, a node is of type lecturer, etc.

Blank Nodes

- Blank nodes do not have a URI and cannot be referenced
  - They can only be subjects or objects
- Its semantics are yet not clear, though
  - De facto use (i.e., most spread use, also in SPARQL): An **identifier** without a URI
  - W3C position: Incomplete data (potentially two blank nodes might be the same resource)
    - An unknown value,
    - A value that does not apply / anonymized value
- The de facto use is a pragmatic use
  - Facilitates reasoning (CWA)
- The Linked Open Data community discourages its use. Everything should have a URI!

Blank Nodes: Reification (I)

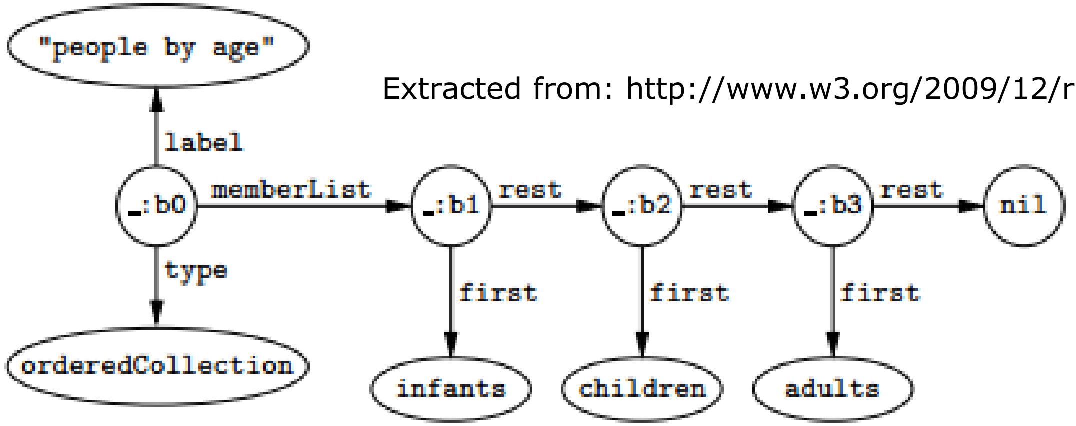
- Example of use ("quoting"):



- In this example, what is schema and what instances?

Blank Nodes: Reification (II)

- Example of use (container):



Extracted from: http://www.w3.org/2009/12/rdf-ws/papers/ws23

Fig. 2. An ordered collection in SKOS.

- These statements can be written as:
  - [*property object*] (e.g., [*label "people by age"*])
    - The subject is considered to be a blank node

RDF-star (RDF-star)

- RDF-star is an RDF extension to elegantly solve reification
  - Compact and precise syntax for reification

    https://w3c.github.io/rdf-star/cg-spec/2021-02-18.html
- Example:

```
@prefix :    <http://www.example.org/> .

:employee38 :familyName "Smith" .
:employee22 :claims << :employee38 :jobTitle "Assistant Designer" >> .
```

Embedded triple

- SPARQL-star is an extension of SPARQL to query RDF-star

Summary

- Sharing data for integration od data exchange purposes require new modeling standards
- 5-stars Open Data is often called Semantic Modeling as it defines the meaning of the data by pre-defined URIs that can be universally located and dereferenced
- RDF
  - Modeling in RDF
  - Blank Nodes

Bibliography

- S. Abiteboul et al. Web Data Management, 2012 (chapter 7, until section 7.3.3)
- Ian Robinson et al. Graph Databases. O'Reilly. 2013 (http://graphdatabases.com/)
- RDF. W3C Recommendation. Latest version at http://www.w3.org/TR/rdf-concepts/