Shaping Knowledge Graphs
ISWC'24 Tutorial

Jose Emilio Labra Gayo
WESO Research group
University of Oviedo, Spain

About me…

Main researcher at WESO (Web Semantics Oviedo)

Some books:

- "*Web semántica*" (in Spanish), 2012
- "*Validating RDF data*", 2017
- "Knowledge Graphs", 2021

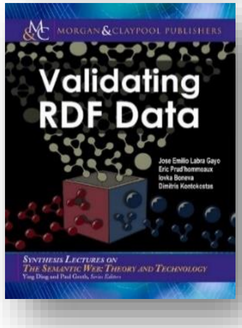
…and software:

- SHaclEX (Scala library, implements ShEx & SHACL)
- RDFShape (RDF playground)
- rudof (RDF & Shapes library in Rust)

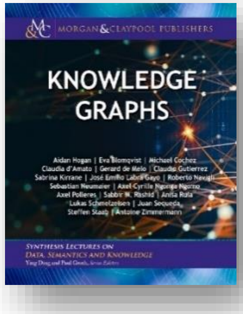
RUDOF

http://labra.weso.es

2012



2017 HTML version:
http://book.validatingrdf.com



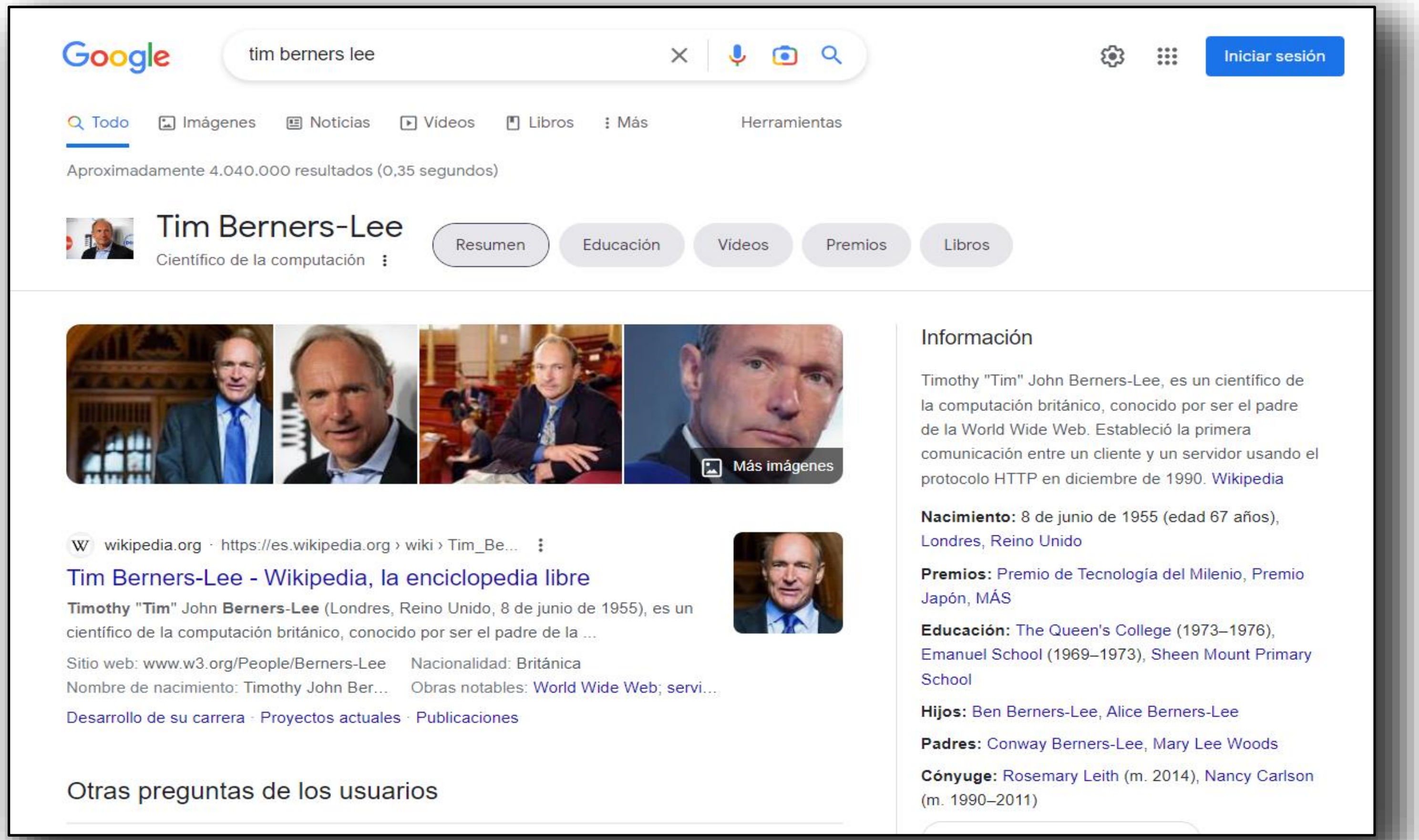
2021, HTML version
https://kgbook.org/

Contents

- Introduction to Knowledge graphs
- Types of Knowledge Graphs:
    - RDF, Property graphs, Wikibase, RDF-Star
- Validating RDF: ShEx & SHACL
- Validating Property Graphs
- Validating Wikibase and Wikidata graphs
- Validating RDF-Star
- Applications:
    - Inferring shapes from data, Knowledge Graphs Subsets, etc.

Knowledge Graphs

Current notion of Knowledge Graphs Popular after Google, 2012*



Aproximadamente 4.040.000 resultados (0,35 segundos)

Tim Berners-Lee
Científico de la computación

wikipedia.org · https://es.wikipedia.org › wiki › Tim_Be...

Tim Berners-Lee - Wikipedia, la enciclopedia libre

Timothy "**Tim**" John **Berners**-**Lee** (Londres, Reino Unido, 8 de junio de 1955), es un científico de la computación británico, conocido por ser el padre de la ...

Sitio web: www.w3.org/People/Berners-Lee
Nacionalidad: Británica
Nombre de nacimiento: Timothy John Ber...
Obras notables: World Wide Web; servi...

Desarrollo de su carrera · Proyectos actuales · Publicaciones

Otras preguntas de los usuarios

Información

Timothy "Tim" John Berners-Lee, es un científico de la computación británico, conocido por ser el padre de la World Wide Web. Estableció la primera comunicación entre un cliente y un servidor usando el protocolo HTTP en diciembre de 1990. Wikipedia

Nacimiento: 8 de junio de 1955 (edad 67 años), Londres, Reino Unido

Premios: Premio de Tecnología del Milenio, Premio Japón, MÁS

Educación: The Queen's College (1973–1976), Emanuel School (1969–1973), Sheen Mount Primary School

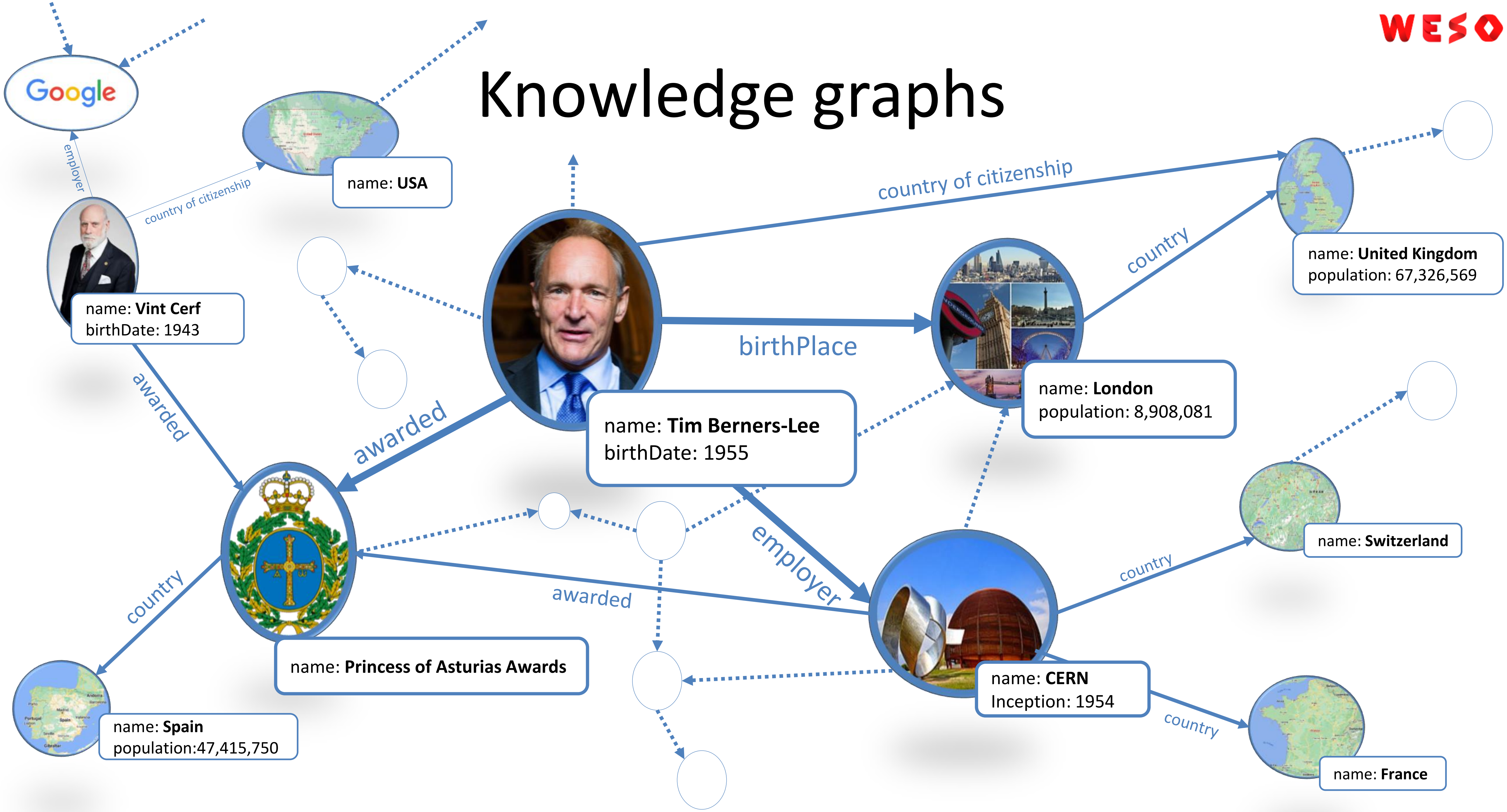
Hijos: Ben Berners-Lee, Alice Berners-Lee

Padres: Conway Berners-Lee, Mary Lee Woods

Cónyuge: Rosemary Leith (m. 2014), Nancy Carlson (m. 1990–2011)

Link: https://www.blog.google/products/search/introducing-knowledge-graph-things-not/

Knowledge graphs



name: **Tim Berners-Lee**
birthDate: 1955

name: **Vint Cerf**
birthDate: 1943

name: **USA**

name: **United Kingdom**
population: 67,326,569

name: **London**
population: 8,908,081

name: **Switzerland**

name: **CERN**
Inception: 1954

name: **France**

name: **Princess of Asturias Awards**

name: **Spain**
population:47,415,750

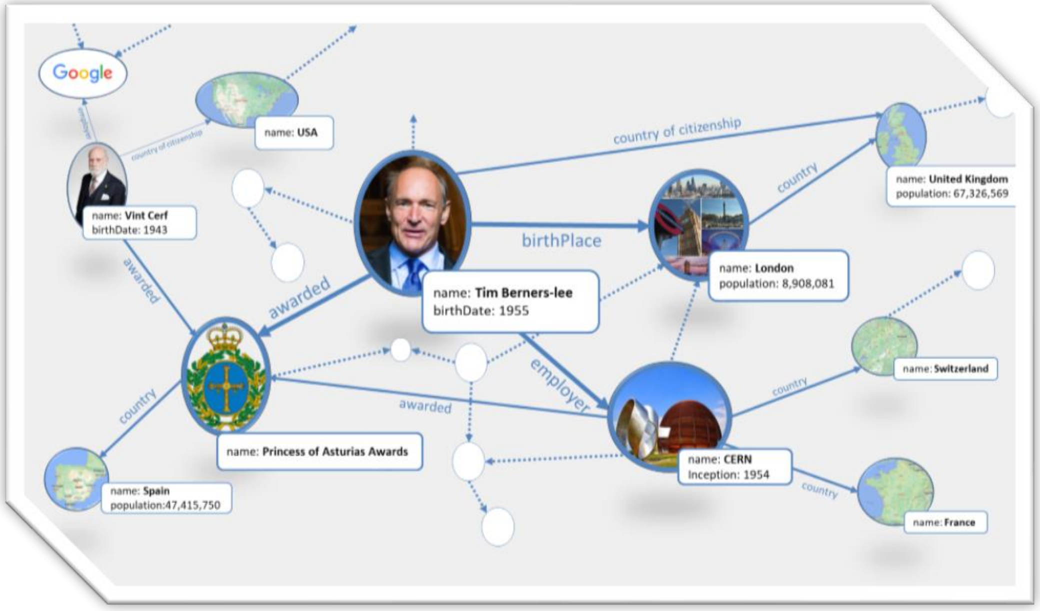
Knowledge graphs

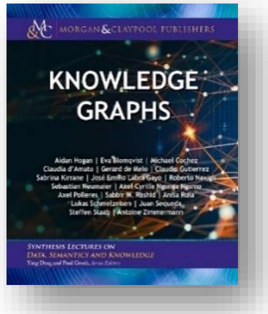
Knowledge graph = *a **graph of data***

*intended to accumulate and convey **knowledge** of the real world*

*whose nodes represent **entities** of interest and*

*whose edges represent **relations** between these entities.*




MORGAN & CLAYPOOL PUBLISHERS
KNOWLEDGE GRAPHS
Aidan Hogan | Eva Blomqvist | Michael Cochez
Claudia d'Amato | Gerard de Melo | Claudio Gutierrez
Sabrina Kirrane | José Emilio Labra Gayo | Roberto Navigli
Sebastian Neumaier | Axel-Cyrille Ngonga Ngomo
Axel Polleres | Sabbir M. Rashid | Anisa Rula
Lukas Schmelzeisen | Juan Sequeda
Steffen Staab | Antoine Zimmermann
SYNTHESIS LECTURES ON DATA, SEMANTICS AND KNOWLEDGE

https://kgbook.org/

Applications of Knowledge graphs

Improve search results

Question answering

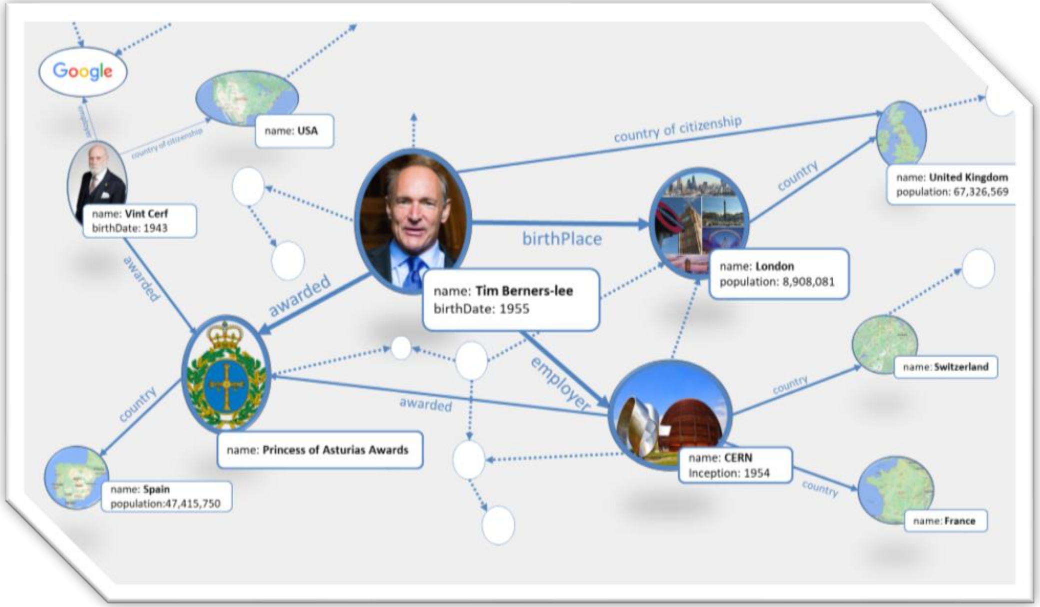
Data governance

Handling heterogenous data

Recommender systems

Chatbots and NLP

. . .



Contents

Introduction to Knowledge graphs

Types of Knowledge Graphs:

RDF, Property graphs, Wikibase, RDF-Star

Validating RDF: ShEx & SHACL

Validating Property Graphs

Validating Wikibase and Wikidata graphs

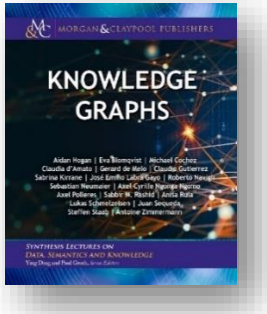
Validating RDF-Star

Applications:

Inferring shapes from data, Knowledge Graphs Subsets, etc.

Types of Knowledge graphs

- Open Knowledge graphs
  - Cross-domain: Wikidata, Dbpedia, Freebase, YAGO, ...
  - Domain specific
    - Academic: Open citations, SciGraph, Microsoft Academic Knowledge Graph, ...
    - Life sciences: UniProt, PubChem, PDB, ...
    - Government: EU Knowledge graph, ...
    - ...
- Enterprise Knowledge graphs
  - Web search: Google, Bing...
  - Commerce: AirBnb, Amazon, eBay, Uber,...
  - Social networks: Linkedin, Facebook,...
  - Finance: Banca d'Italia, Bloomberg, Wells Fargo, Capital One,...
  - ...


MORGAN & CLAYPOOL PUBLISHERS
KNOWLEDGE GRAPHS
Aidan Hogan | Eva Blomqvist | Michael Cochez
Claudia d'Amato | Gerard de Melo | Claudio Gutierrez
Sabrina Kirrane | José Emilio Labra Gayo | Roberto Navigli
Sebastian Neumaier | Axel-Cyrille Ngonga Ngomo
Axel Polleres | Sabbir M. Rashid | Anisa Rula
Lukas Schmelzeisen | Juan Sequeda
Steffen Staab | Antoine Zimmermann
SYNTHESIS LECTURES ON DATA, SEMANTICS AND KNOWLEDGE

https://kgbook.org/

Wikidata as an example

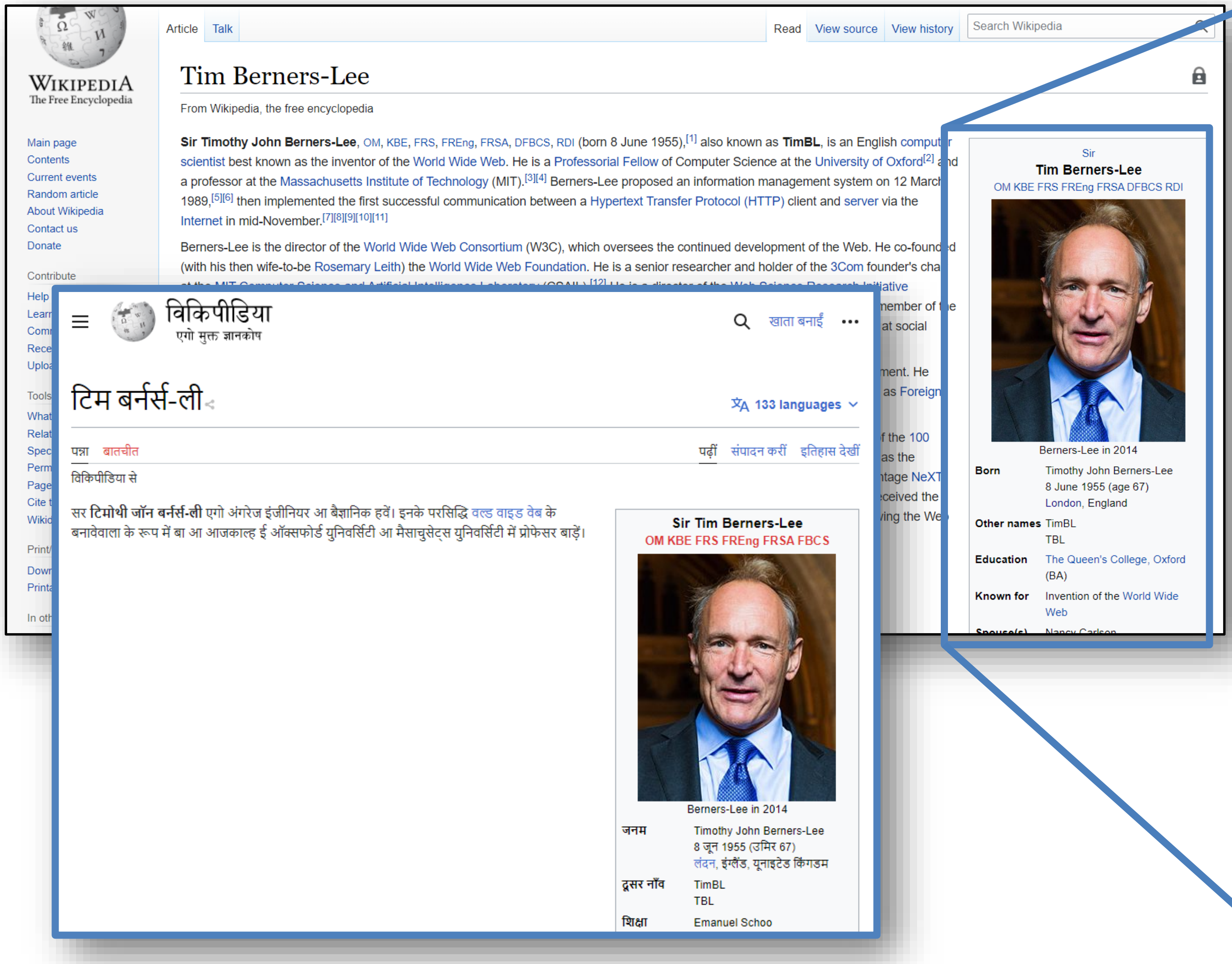
Wikidata created in 2012 as a collaborative knowledge graph

Initial goal:

Support multilingual infoboxes in Wikipedia

Wikidata as an example

English Wikipedia page of Tim Berners-Lee



Tim Berners-Lee

From Wikipedia, the free encyclopedia

Sir Timothy John Berners-Lee, OM, KBE, FRS, FREng, FRSA, DFBCS, RDI (born 8 June 1955),[1] also known as **TimBL**, is an English comput r scientist best known as the inventor of the World Wide Web. He is a Professorial Fellow of Computer Science at the University of Oxford[2] a d a professor at the Massachusetts Institute of Technology (MIT).[3][4] Berners-Lee proposed an information management system on 12 March 1989,[5][6] then implemented the first successful communication between a Hypertext Transfer Protocol (HTTP) client and server via the Internet in mid-November.[7][8][9][10][11]

Berners-Lee is the director of the World Wide Web Consortium (W3C), which oversees the continued development of the Web. He co-found d (with his then wife-to-be Rosemary Leith) the World Wide Web Foundation. He is a senior researcher and holder of the 3Com founder's cha

Sir
Tim Berners-Lee
OM KBE FRS FREng FRSA DFBCS RDI

Berners-Lee in 2014

Born	Timothy John Berners-Lee 8 June 1955 (age 67) London, England
Other names	TimBL TBL
Education	The Queen's College, Oxford (BA)
Known for	Invention of the World Wide Web
Spouse(s)	Nancy Carlson

टिम बर्नर्स-ली

विकिपीडिया से

सर **टिमोथी जॉन बर्नर्स-ली** एगो अंगरेज इंजीनियर आ बैज्ञानिक हवें। इनके परसिद्धि वल्ड वाइड वेब के बनावेवाला के रूप में बा आ आजकाल्ह ई ऑक्सफोर्ड युनिवर्सिटी आ मैसाचुसेट्स युनिवर्सिटी में प्रोफेसर बाड़ें।

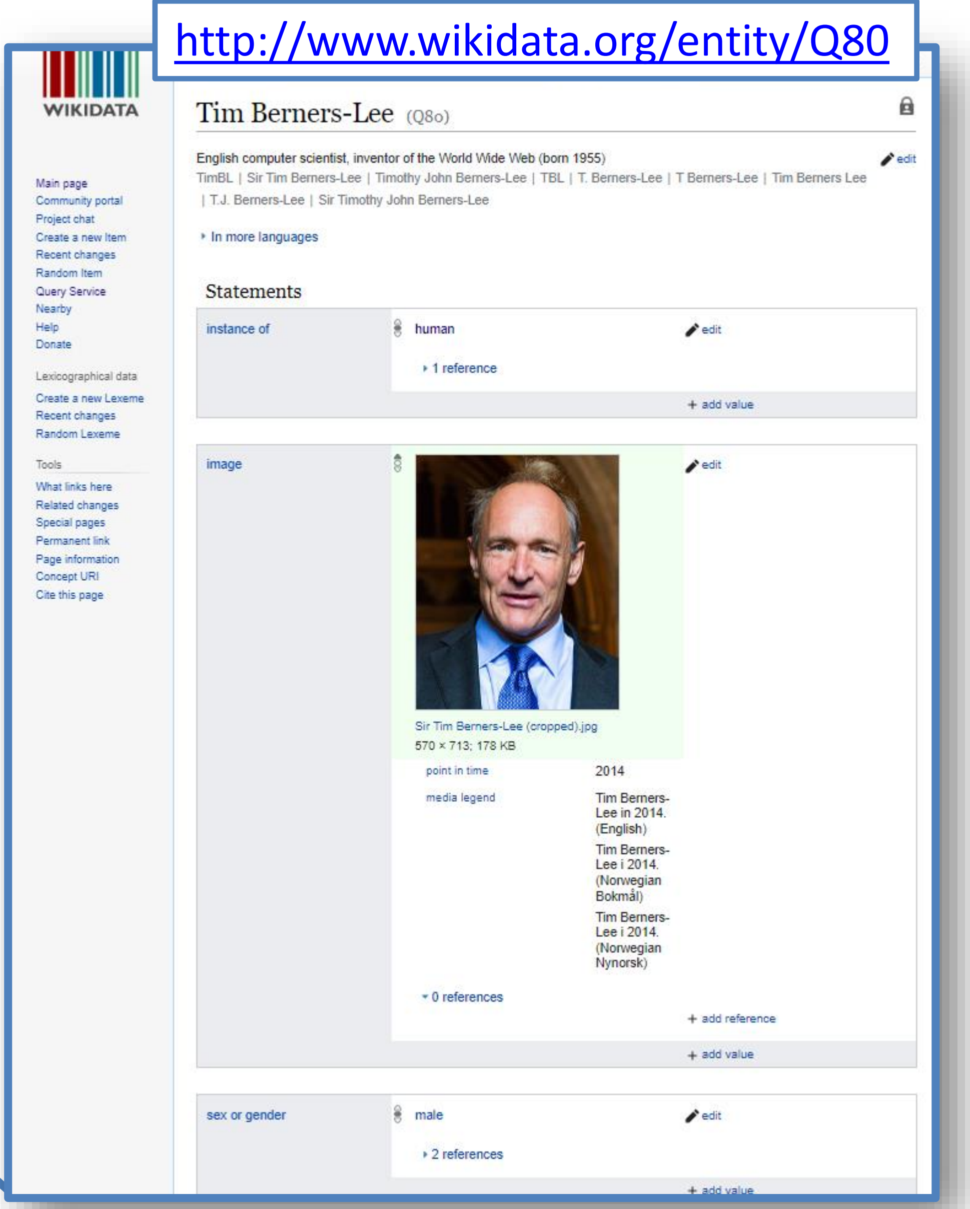
Sir Tim Berners-Lee
OM KBE FRS FREng FRSA FBCS

Berners-Lee in 2014

जनम	Timothy John Berners-Lee 8 जून 1955 (उमिर 67) लंदन, इंग्लैंड, यूनाइटेड किंगडम
दूसर नाँव	TimBL TBL
शिक्षा	Emanuel Schoo

Bihari Wikipedia of Tim Berners-Lee

http://www.wikidata.org/entity/Q80



Tim Berners-Lee (Q80)

English computer scientist, inventor of the World Wide Web (born 1955)

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | T. Berners-Lee | T Berners-Lee | Tim Berners Lee | T.J. Berners-Lee | Sir Timothy John Berners-Lee

Statements

instance of: human
1 reference

image: Sir Tim Berners-Lee (cropped).jpg
570 × 713; 178 KB
point in time: 2014
media legend: Tim Berners-Lee in 2014. (English)
Tim Berners-Lee i 2014. (Norwegian Bokmål)
Tim Berners-Lee i 2014. (Norwegian Nynorsk)
0 references

sex or gender: male
2 references

Wikidata: some features

Collaborative: anyone can edit

Free and open license

Currently (01/2023): 101m items, 1,8b edits

Co-edited by humans and bots: 23k active users, 343 bots

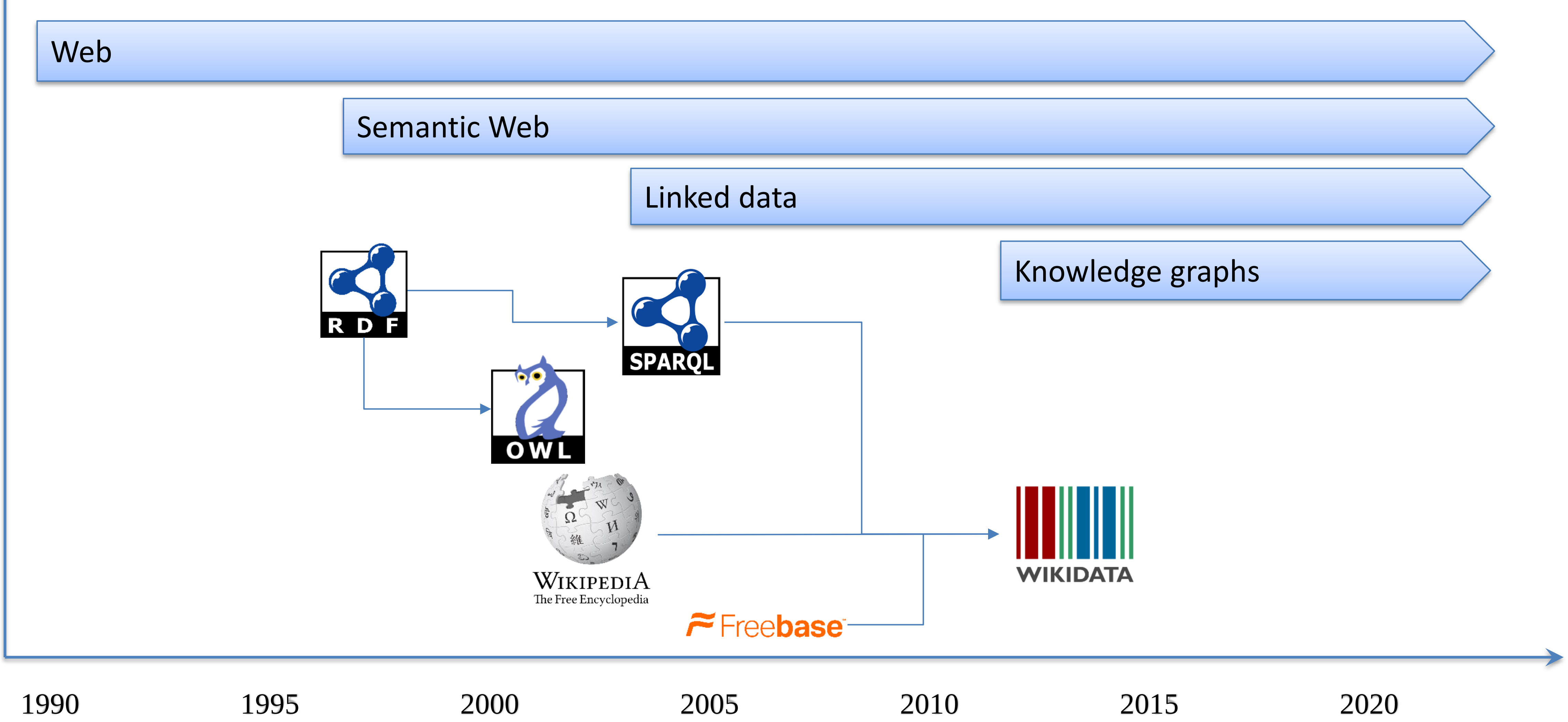
Open Wikidata Query Service: Public SPARQL endpoint

Dumps freely available: 109Gb compressed

Software that supports Wikidata = Wikibase

Evolution

Timeline with some concepts and technologies...



Web

Semantic Web

Linked data

Knowledge graphs

Knowledge Graphs models

3 popular knowledge graphs models

- RDF based
- Property graphs
- Wikibase graphs

RDF graphs

RDF = W3C recommendation (since 98)

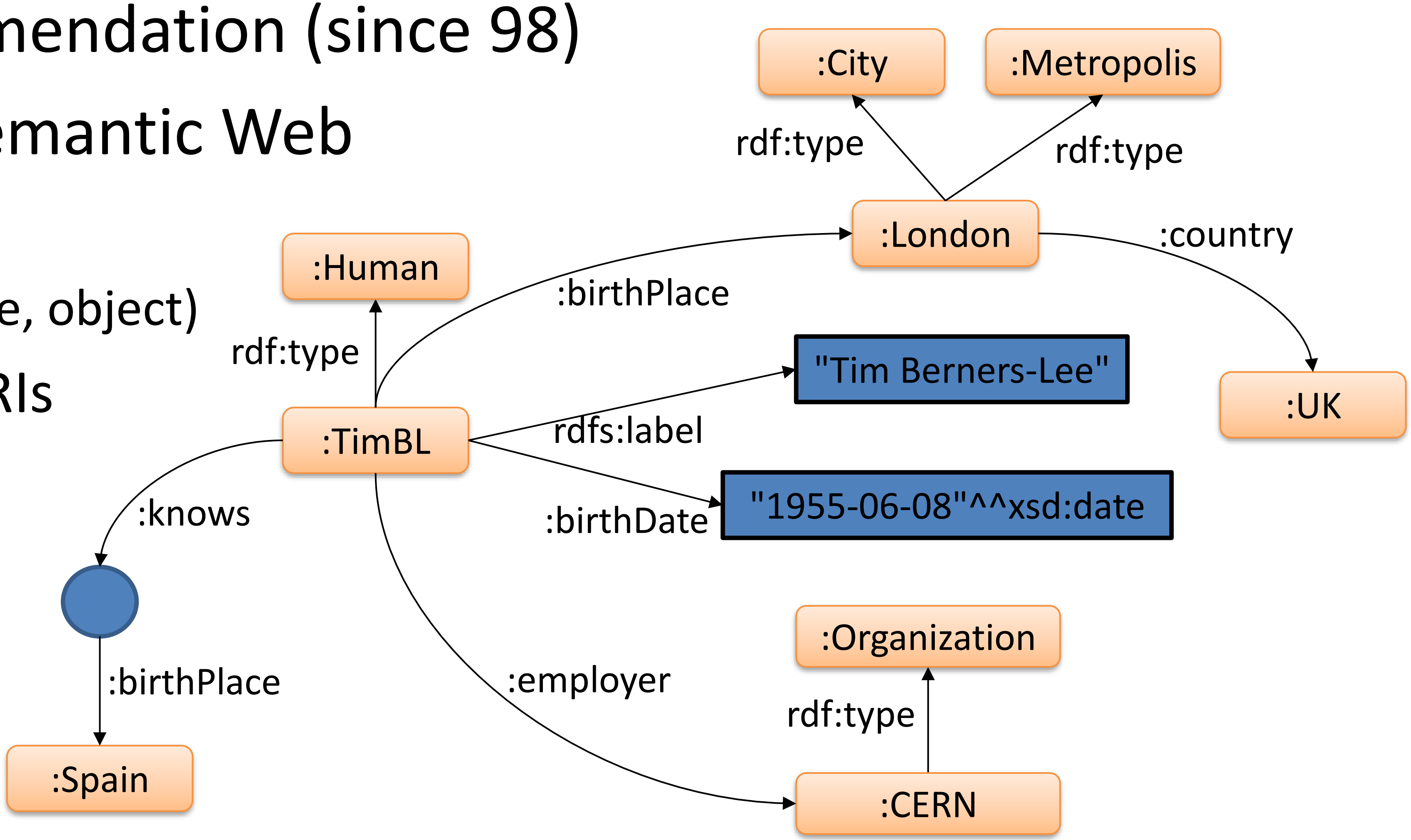
Lingua franca of Semantic Web

Based on triples

(subject, predicaje, object)

Most nodes are URIs

Interoperability



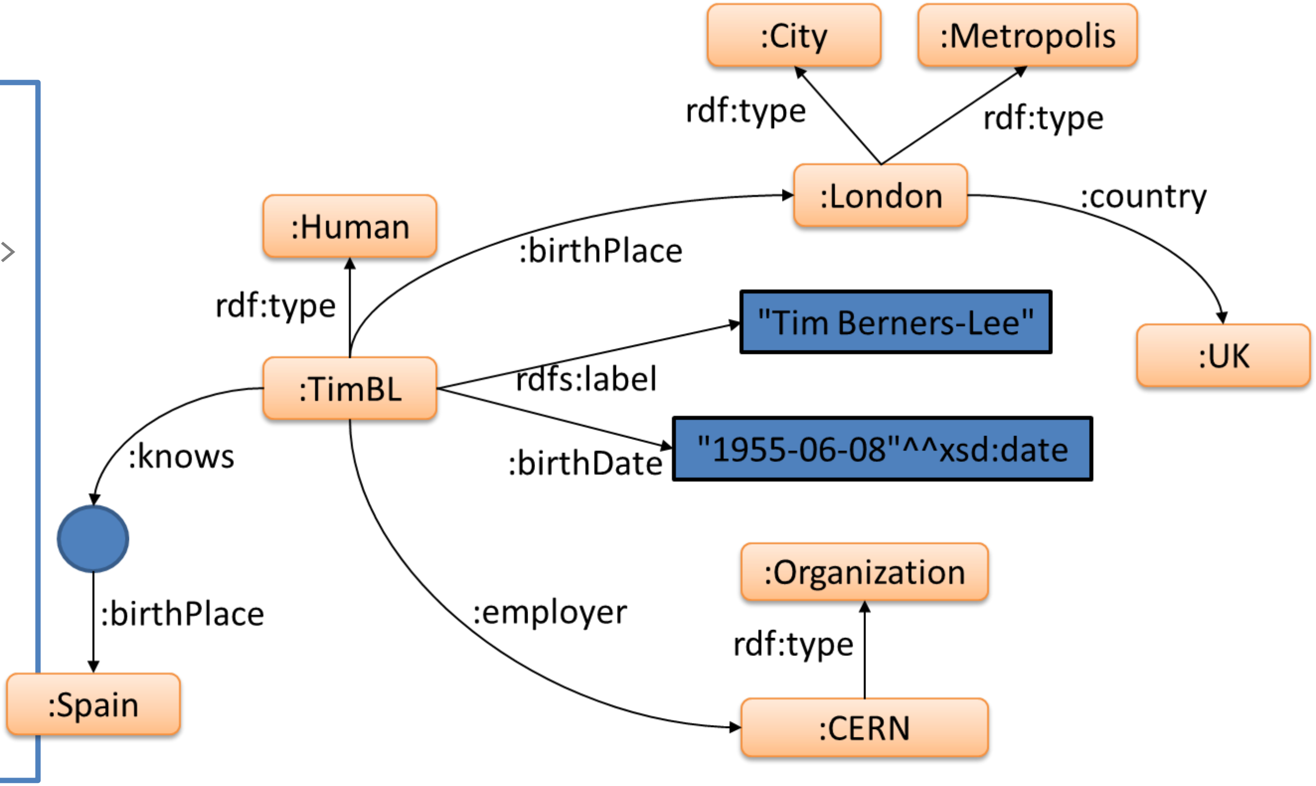
RDF ecosystem

One data model, several syntaxes: Turtle, N-Triples, JSON-LD

Vocabularies: RDF Schema, OWL, SKOS, etc.

Turtle

```
prefix :      <http://example.org/>
prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
prefix xsd:   <http://www.w3.org/2001/XMLSchema#>
prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

:timbl  rdf:type    :Human ;
        :birthPlace :london ;
        rdfs:label  "Tim Berners-Lee" ;
        :birthDate  "1955-06-08"^^xsd:date ;
        :employer   :CERN ;
        :knows      _:1   .
:london rdf:type    :City, :Metropolis ;
        :country    :UK .
:CERN   rdf:type    :Organization .
_:1     :birthPlace :Spain .
```



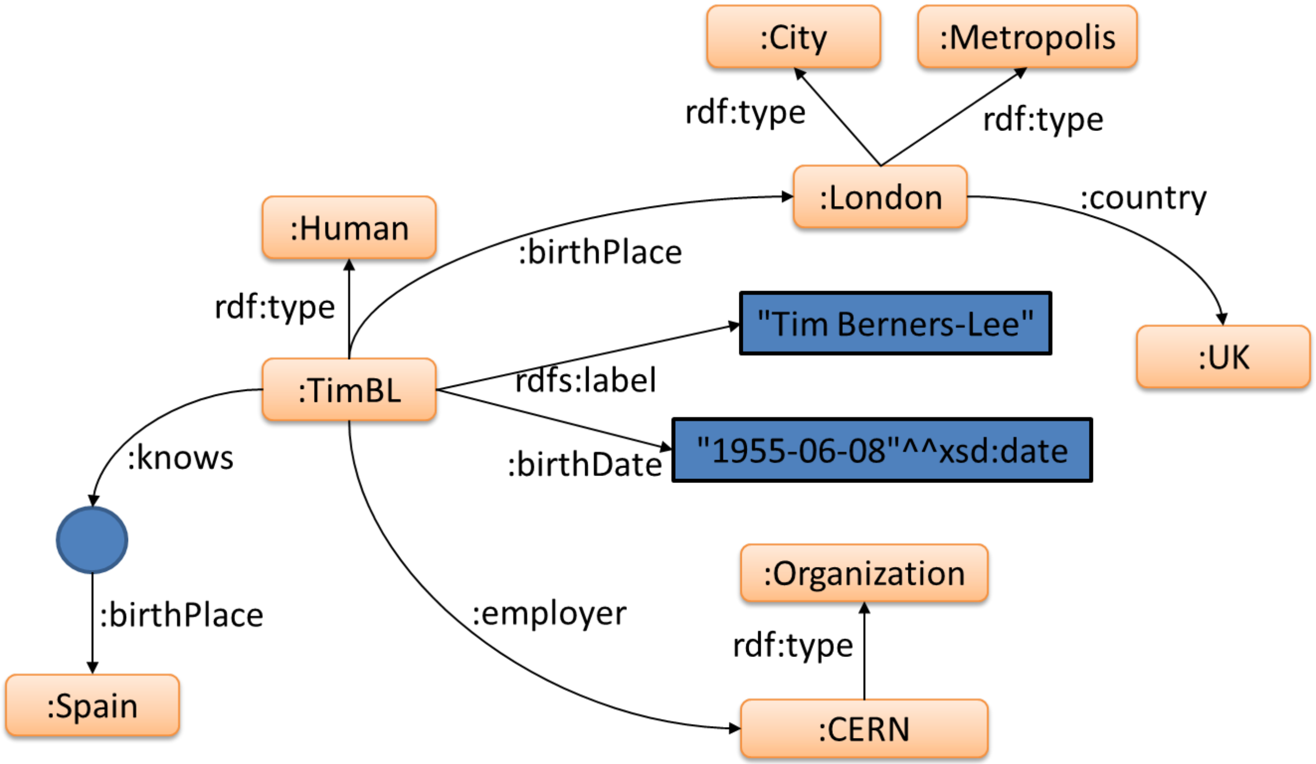
RDF ecosystem: SPARQL

SPARQL is an RDF query language and protocol

It enables the creation of SPARQL endpoints

```
select ?person ?date ?country where {
    ?person :birthDate  ?date .
    ?person :birthPlace ?p .
    ?p      :country    ?country
}
```

?person	?date	?country
:timbl	1955-06-08	:UK



RDF1.2 (RDF-Star)

Triple terms

Add statements about triples

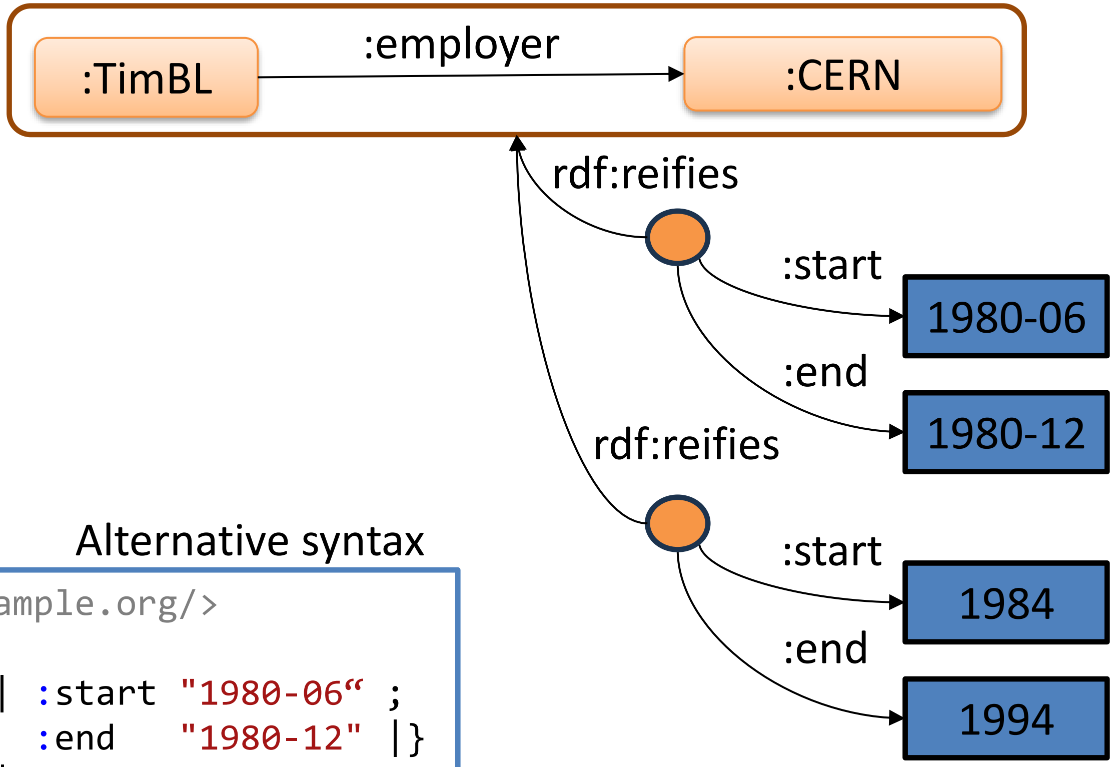
Reifiers

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

_:r1 rdf:reifies << :timbl :employer :CERN >> .
_:r1 :start "1980-06";
     :end   "1980-12" .
```

```
_:r1 rdf:reifies << :timbl :employer :CERN >> .
_:r2 :start "1984;
     :end   "1994" .
```

Alternative syntax

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

:timbl :employer :CERN {| :start "1980-06“ ;
                          :end   "1980-12" |}
                       {| :start "1984“ ;
                          :end   "1994" |} .
```



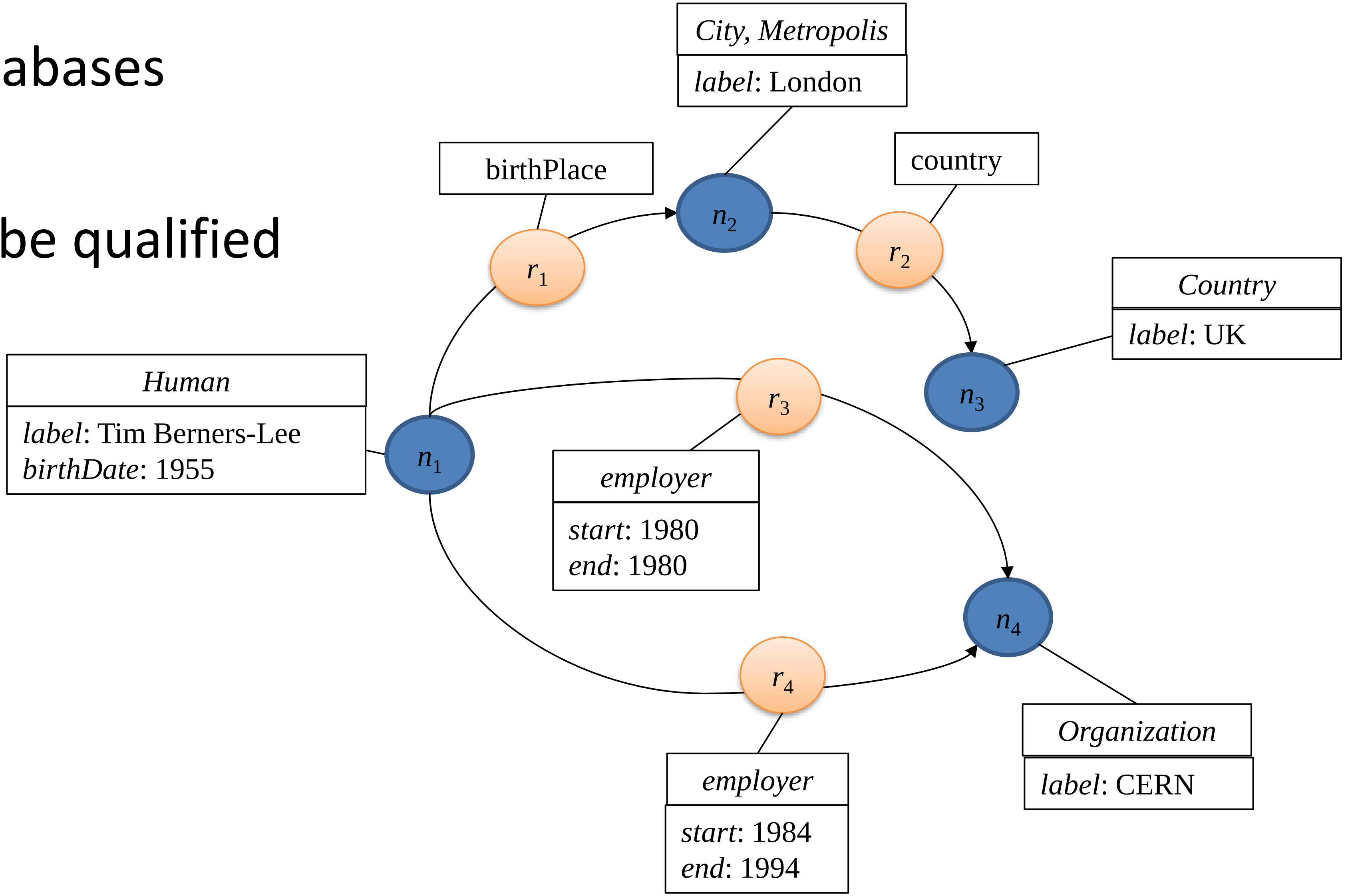
Property graphs

Popularized by graph databases

Example: Neo4j

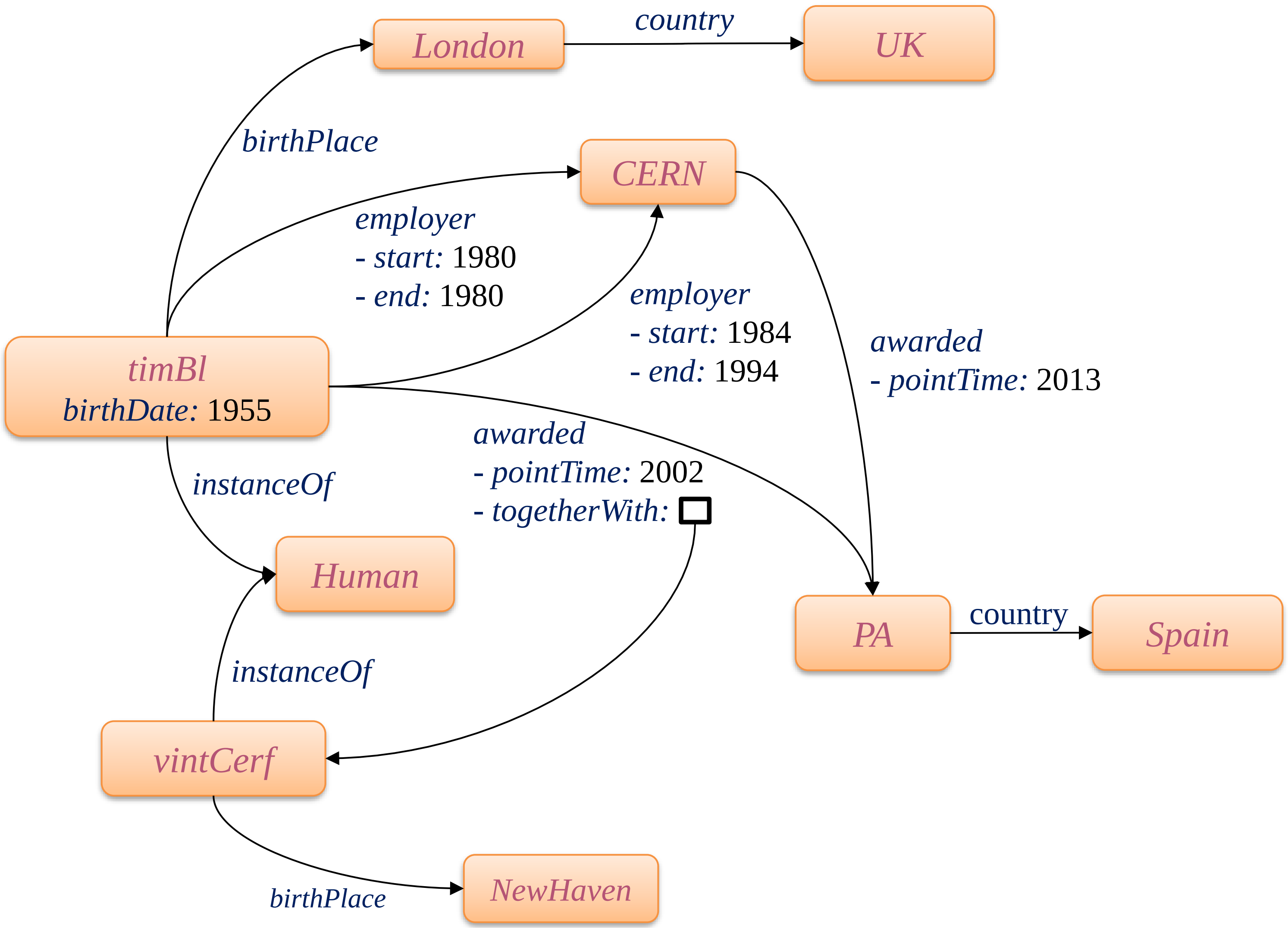
Nodes and relations can be qualified

Nodes can have types



Wikibase graphs

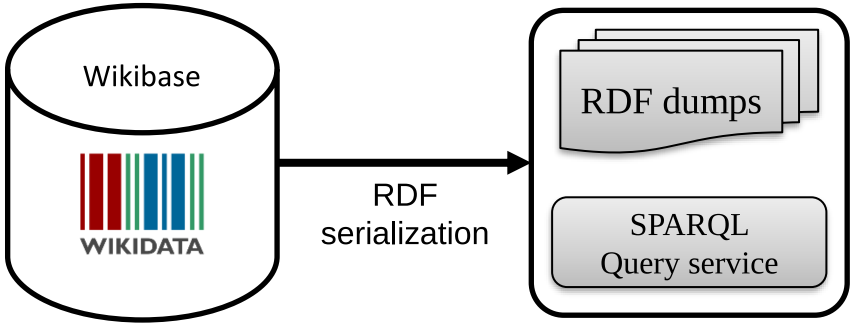
Popularized by Wikidata
Wikibase = software supporting Wikidata
The values can be nodes in the graph
Example:
Tim Berners Lee
http://www.wikidata.org/entity/Q80



Wikibase graphs and SPARQL

Wikibase graphs generate RDF serializations for each item

SPARQL endpoint and Query service available



```
select ?name ?date?country where {
 wd:Q80 wdt:P1559 ?name .
 wd:Q80 wdt:P569  ?date .
 wd:Q80 wdt:P19   ?place .
 ?place wdt:P17   ?country
}
```

?name	?date	?country
Tim Berners-lee	1955-06-08	:UK

Try it: https://w.wiki/5yGu

Contents

Introduction to Knowledge graphs

Types of Knowledge Graphs:

- RDF, Property graphs, Wikibase, RDF-Star

👉 Validating RDF: ShEx & SHACL

Validating Property Graphs

Validating Wikibase and Wikidata graphs

Validating RDF-Star

Applications:

- Inferring shapes from data, Knowledge Graphs Subsets, etc.

RDF Data Model

Overview of RDF Data Model and simple exercise

Link to slides about
RDF Data Model

https://doi.org/10.6084/m9.figshare.13174562

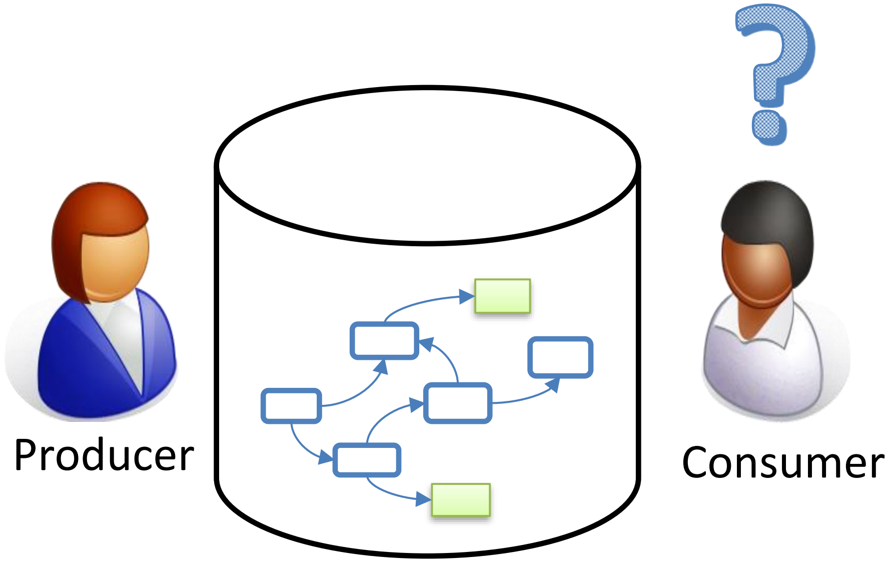
RDF, the good parts...

- RDF as an integration language
- RDF as a *lingua franca* for semantic web and linked data
- RDF flexibility
  - Data can be adapted to multiple environments
  - Open and reusable data by default
- RDF for knowledge representation
- RDF data stores & SPARQL

RDF, the other parts

Consuming & producing RDF

- Multiple serializations: Turtle, RDF/XML, JSON-LD, ...
- Embedding RDF in HTML
- Describing and validating RDF content



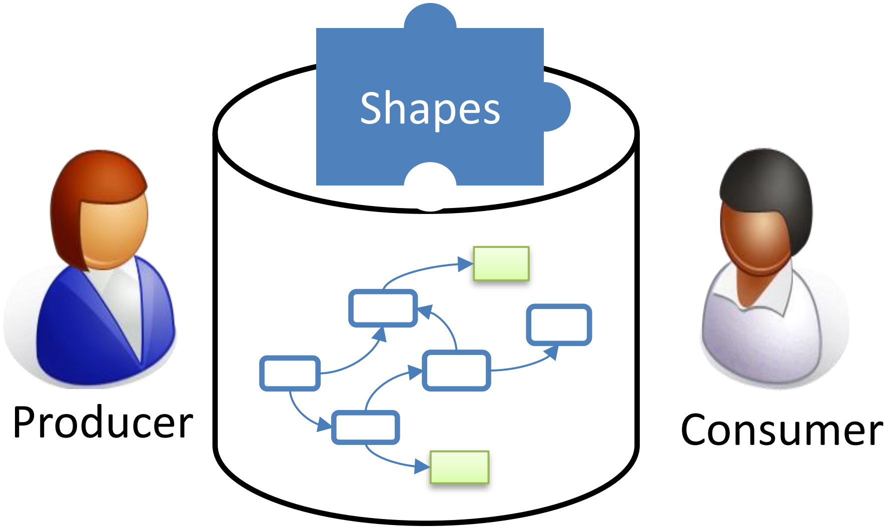
Why describe & validate RDF?

For producers

- Developers can understand the contents they are going to produce
- They can ensure they produce the expected structure
- Advertise and document the structure
- Generate interfaces

For consumers

- Understand the contents
- Verify the structure before processing it
- Query generation & optimization



Similar technologies

Technology	Schema
Relational Databases	DDL
XML	DTD, XML Schema, RelaxNG, Schematron
Json	Json Schema
RDF	?

Fill that gap

Understanding the problem

RDF is composed by nodes and arcs between nodes

We can describe/check

- The form of the node itself (node constraint)
- The number of possible arcs incoming/outgoing from a node
- The possible values associated with those arcs

RDF Node

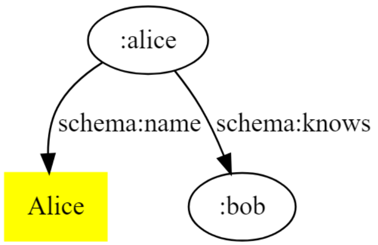
```
:alice schema:name   "Alice";
       schema:knows  :bob  .
```

Abstract shape of a node that represents a User

```
IRI schema:name   string   1
    schema:knows  IRI      0, 1,...
```

ShEx

```
<UserShape> IRI {
  schema:name   xsd:string    ;
  schema:knows  IRI         *
}
```

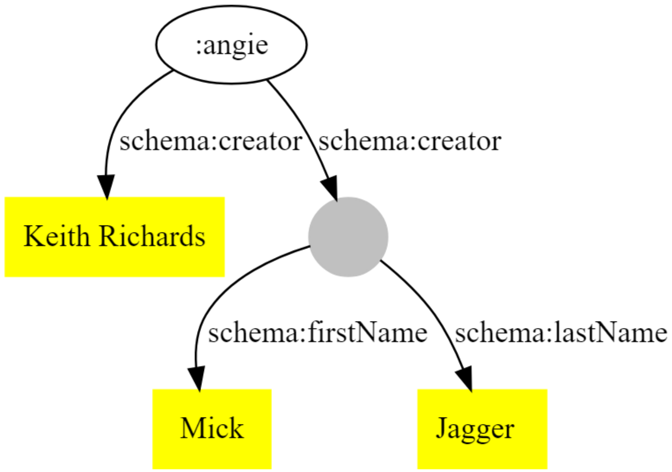


Understanding the problem

RDF flexibility

- Mixed use of objects & literals
- Example:
  - Values of schema:creator can be:
    - string or schema:Person
  - in the same data

```
:angie schema:creator "Keith Richards" ;
       schema:creator [
         schema:firstName "Mick" ;
         schema:lastName  "Jagger"
       ] .
```



Lots of examples at http://schema.org

Understanding the problem

Repeated properties

The same property can be used for different purposes in the same data

Example: A product must have 2 codes with different structure

```
:product schema:productID "isbn:123-456-789";
         schema:productID "code456" .
```

A practical example from FHIR

See: http://hl7-fhir.github.io/observation-example-bloodpressure.ttl.html

Understanding the problem

Shapes ≠ types

Nodes in RDF graphs can have zero, one or many `rdf:type` declarations

- One type can be used for multiple purposes (`schema:Person`)
- Data doesn't need to be annotated with fully discriminating types
    - Nodes with type `schema:Person` can also be customers, patients, etc...
    - Different meanings and different structure depending on the context
    - Different validation constraints at different contexts

Shapes vs Ontology

Ontologies ≠ Shapes ≠ instance data

- Ontologies are usually focused on domain entities (higher level)
- RDF validation/shapes focused on RDF graph features (lower level)

Different levels

Ontology

```
:Person a owl:Class ;
   rdfs:subClassOf [a owl:Restriction ;
                    owl:onProperty  :hasParent ;
                    owl:qualifiedCardinality 2 ;
                    owl:onClass :Person ].
```

Shapes
RDF *Validation*
Constraints

```
<PersonShape> IRI {
 :hasParent @<PersonShape> {0,2}
}
```

Instance data

```
:alice :hasParent :bob, :carol .
:bob   :hasParent  :dave .
```

Previous RDF validation approaches

- SPARQL based
  - Plain SPARQL
  - SPIN: http://spinrdf.org/
- OWL based
  - Stardog ICV
    - http://docs.stardog.com/icv/icv-specification.html
- Grammar based
  - OSLC Resource Shapes
    - https://www.w3.org/Submission/2014/SUBM-shapes-20140211/

Define SPARQL queries that detect errors

Pros:

- Expressive
- Ubiquitous

Cons

- Expressive
- Idiomatic - many ways to encode the same constraint

Example: SPARQL query to check that...

There is one schema:name which must be a xsd:string and one schema:gender must be schema:Male or schema:Female

```
ASK {{ SELECT ?Person {
      ?Person schema:name ?o .
    } GROUP BY ?Person HAVING (COUNT(*)=1)
  }
  { SELECT ?Person {
      ?Person schema:name ?o .
      FILTER ( isLiteral(?o) &&
               datatype(?o) = xsd:string )
     } GROUP BY ?Person HAVING (COUNT(*)=1)
  }
  { SELECT ?Person (COUNT(*) AS ?c1) {
      ?Person schema:gender ?o .
    } GROUP BY ?Person HAVING (COUNT(*)=1)}
    { SELECT ?Person (COUNT(*) AS ?c2) {
      ?S schema:gender ?o .
      FILTER ((?o = schema:Female ||
               ?o = schema:Male))
    } GROUP BY ?Person HAVING (COUNT(*)=1)}
    FILTER (?c1 = ?c2)
}
```

SPIN

- SPARQL inferencing notation http://spinrdf.org/
  - Developed by TopQuadrant
  - Commercial product
- Vocabulary associated with user-defined functions in SPARQL
- SPIN has influenced SHACL (see later)

Stardog ICV

ICV - Integrity Constraint Validation

  Commercial product

OWL with unique name assumption and closed world

Compiled to SPARQL

More info: http://docs.stardog.com/icv/icv-specification.html

OSLC Resource Shapes

OSLC Resource Shapes

https://www.w3.org/Submission/shapes/

- Grammar based approach
- Language for RDF validation
- Input for ShEx and SHACL

```
:user a rs:ResourceShape ;
 rs:property [
  rs:name "name" ;
  rs:propertyDefinition schema:name ;
  rs:valueType xsd:string ;
  rs:occurs rs:Exactly-one ;
] ;
 rs:property [
  rs:name "gender" ;
  rs:propertyDefinition schema:gender ;
  rs:allowedValue schema:Male, schema:Female ;
  rs:occurs rs:Zero-or-one ;
].
```

Other approaches

Dublin Core Application profiles (K. Coyle, T. Baker)

http://dublincore.org/documents/dc-dsp/

RDF Data Descriptions (Fischer et al)

http://ceur-ws.org/Vol-1330/paper-33.pdf

RDFUnit (D. Kontokostas)

http://aksw.org/Projects/RDFUnit.html

…

ShEx and SHACL

- 2013 RDF Validation Workshop
    - Conclusions of the workshop:
        - *There is a need of a higher level, concise language for RDF Validation*
    - ShEx initially proposed (v 1.0)
- 2014 W3c Data Shapes WG chartered
- 2017 SHACL accepted as W3C recommendation
- 2017 ShEx 2.0 released as W3C Community group draft
- 2019 ShEx adopted by Wikidata

Continue this tutorial with...

ShEx by example	https://doi.org/10.6084/m9.figshare.13174571
SHACL by example	https://doi.org/10.6084/m9.figshare.13174577
ShEx and SHACL compared	https://doi.org/10.6084/m9.figshare.13174583
Shapes applications and tools	https://doi.org/10.6084/m9.figshare.13174586