

Advanced Topics in Distributed Systems

Philippe Cudré-Mauroux

September 2013

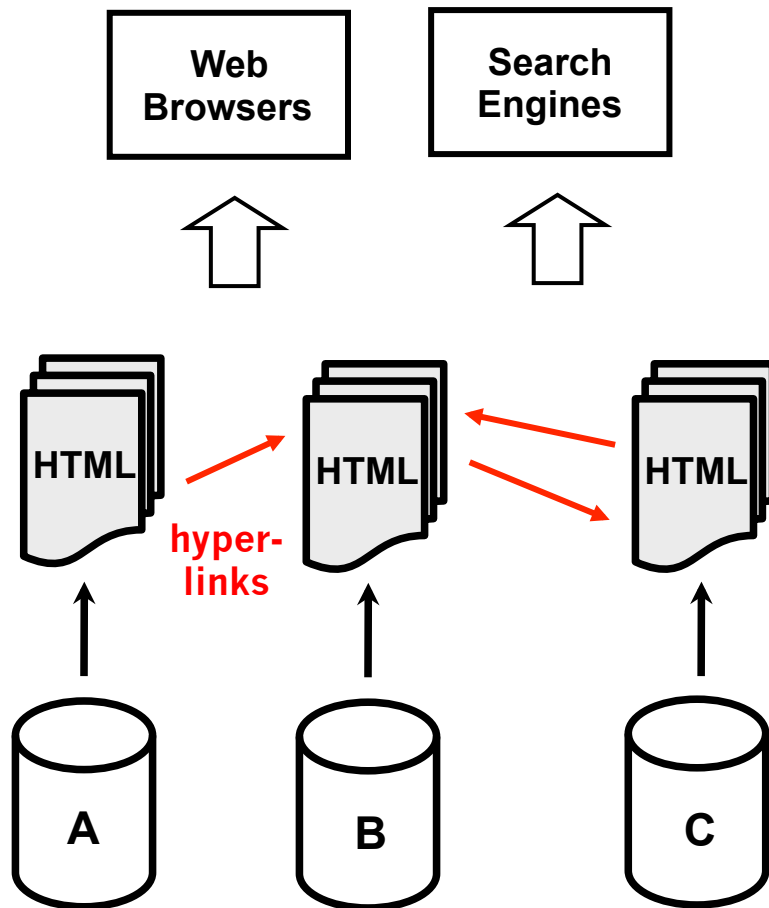
Erasmus Mundus Program, KTH--Sweden

Lecture 4 – The Web of Data

Outline

- Motivation
 - [Web of knowledge, Web of data](#)
- Linked open Data (LoD)
- Facebook Open Graph Protocol
- LoD standards!
 - RDF, RDFS, OWL, SPARQL etc.

The Classic Web



Single Global Information Space Made for Humans

1. URLs as
 - globally unique IDs
 - retrieval mechanism
2. HTML as shared content format
3. Hyperlinks

Problem

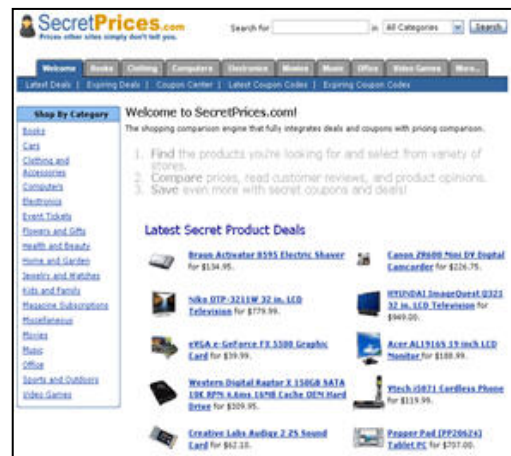
Problem

As Web content is only loosely structured it is difficult for applications to do smart things with it.

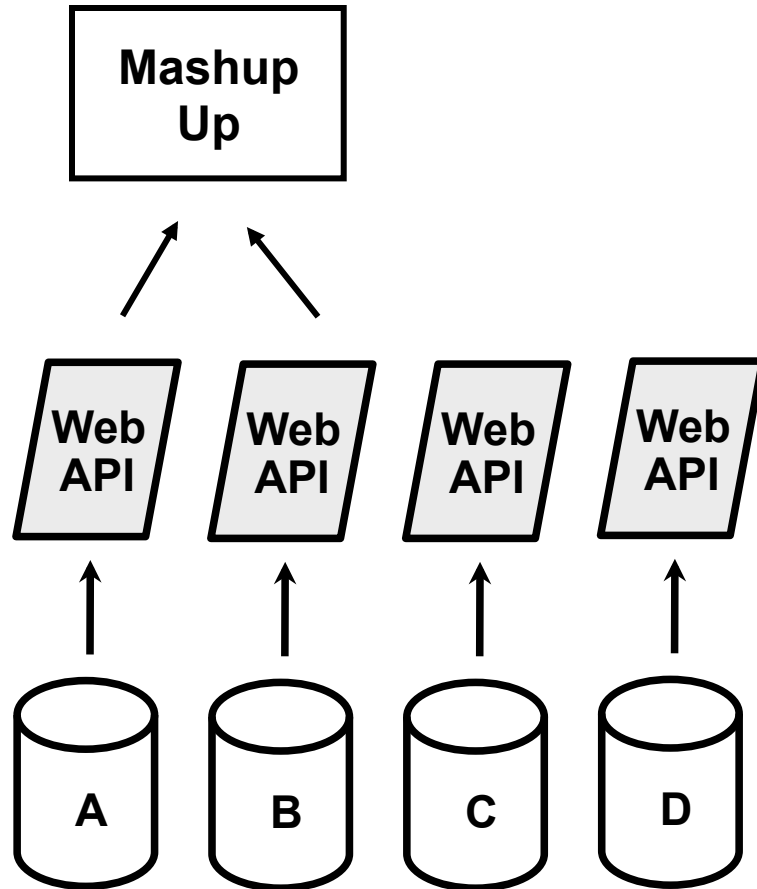
Solution

Increase the structure of Web content.

Web APIs and Mashups



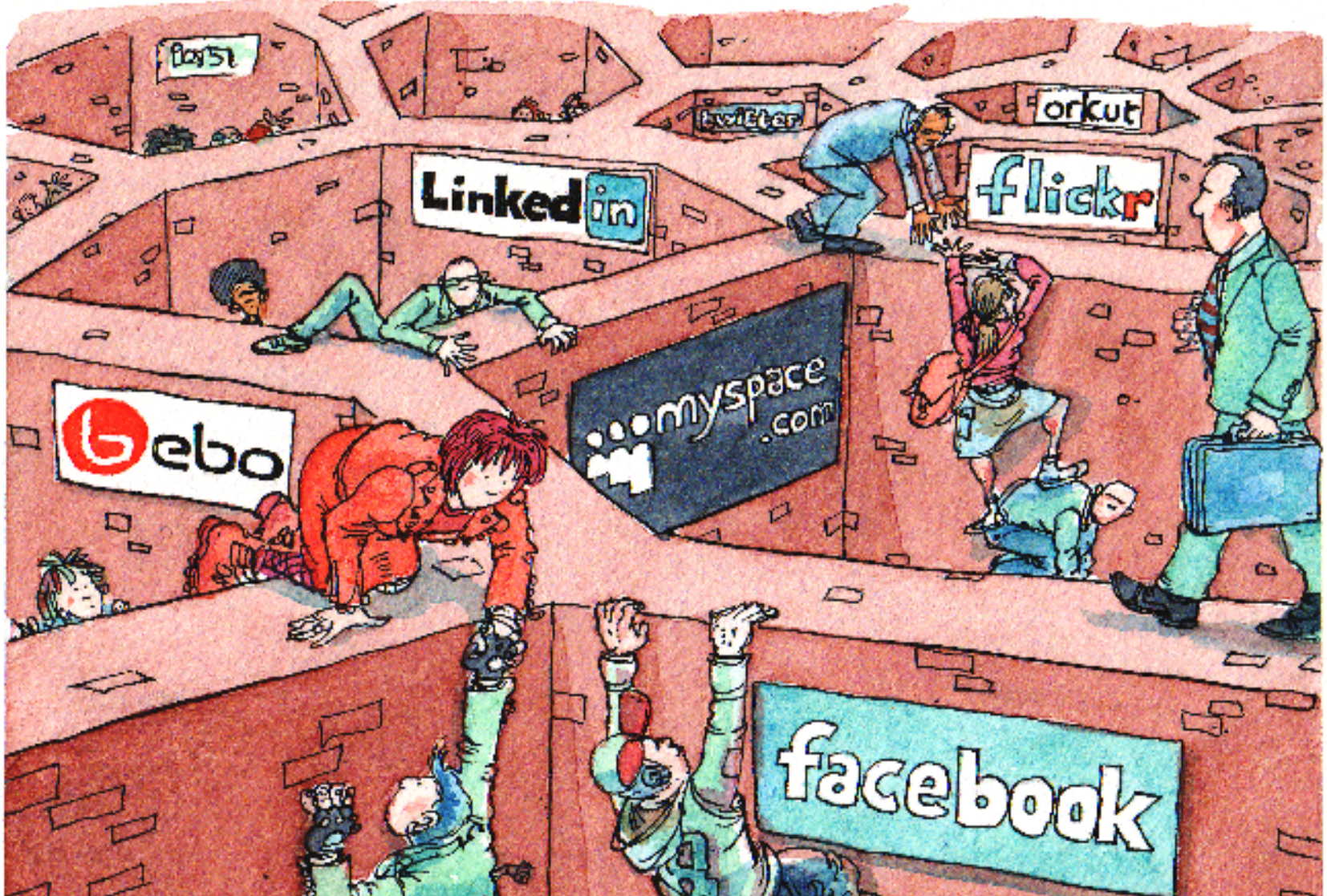
Web APIs and Mashups



Shortcomings

1. APIs provide proprietary interfaces
2. Mashups are based on a fixed set of data sources.
3. You can not set hyperlinks between data objects.

Web APIs slice the Web into Walled Gardens

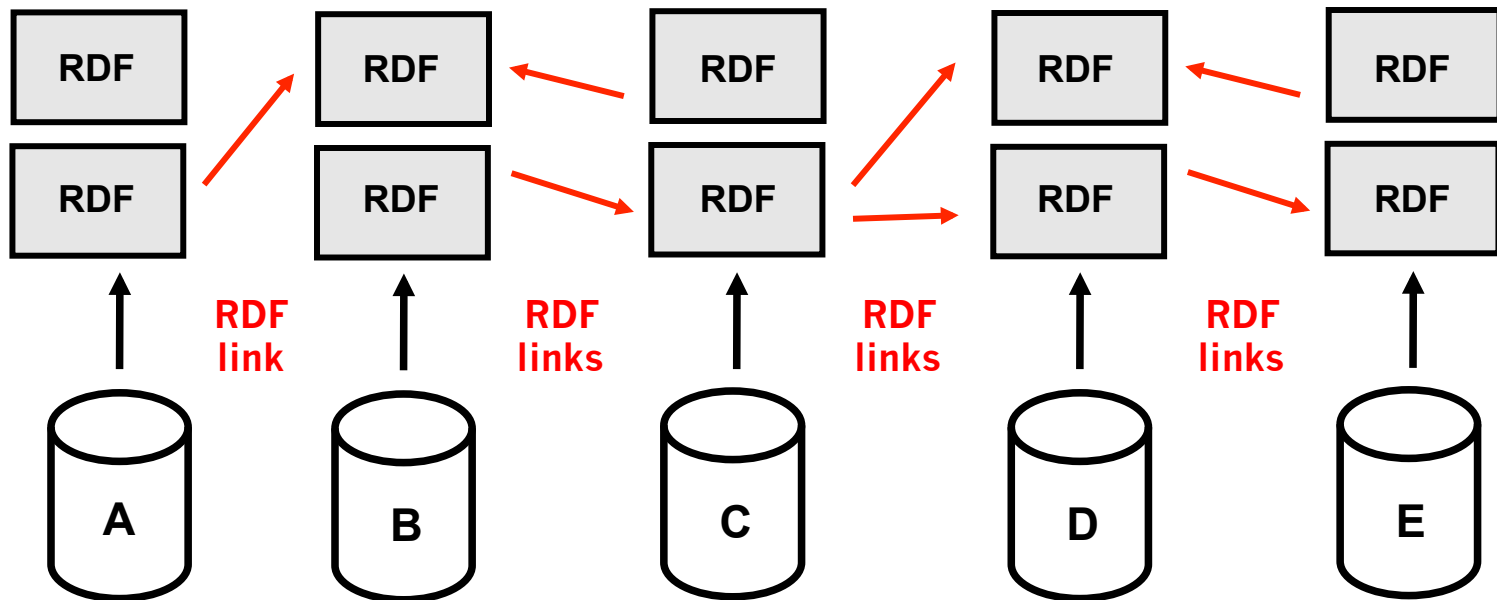


Linked Data



Use Semantic Web technologies to

- 1. publish structured data on the Web,**
- 2. set links between data from one data source to data within other data sources.**



Linked Data Principles



1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names (i.e., use *dereferenceable* URIs).
3. When someone looks up a URI, provide useful structured information, e.g., in RDF
4. Include links to related URIs

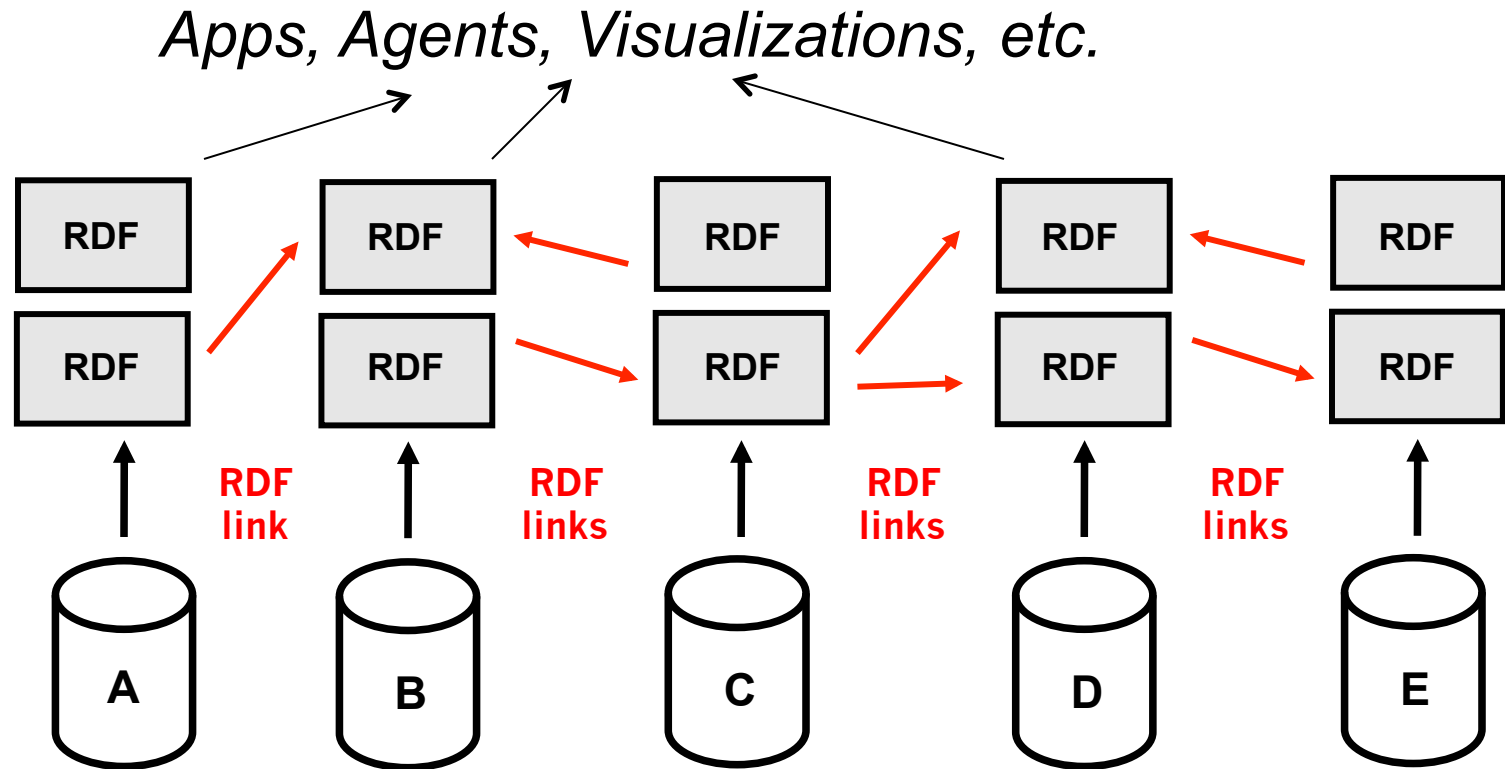
Tim Berners-Lee 2007

<http://www.w3.org/DesignIssues/LinkedData.html>

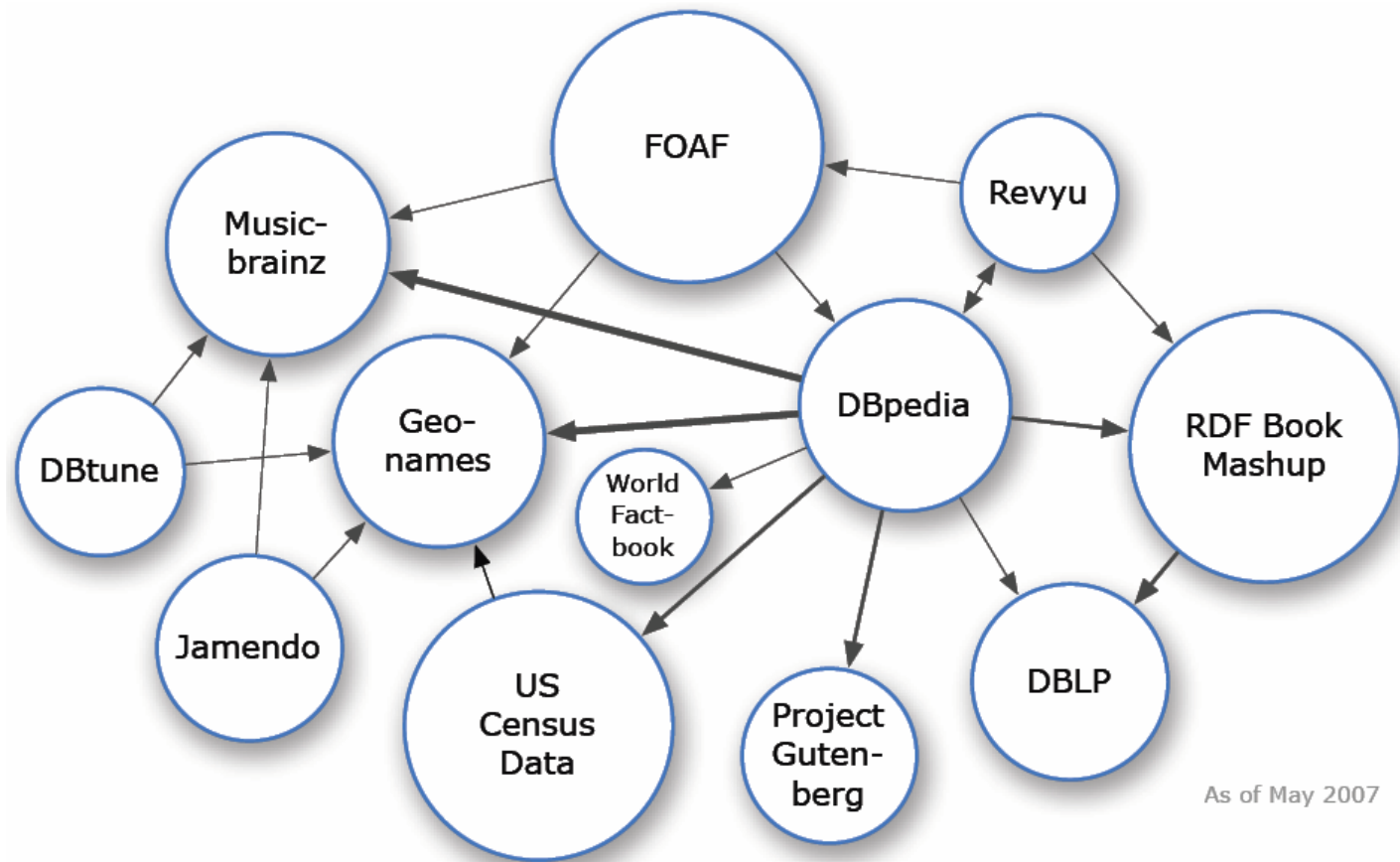
Properties of the Web of Linked Data

- Anyone can publish data to the Web of Linked Data
- Entities are connected by links
 - creating a global data graph that spans data sources and enables the discovery of new data sources.
- Data is self-describing
 - If an application encounters data represented using an unfamiliar vocabulary, the application can resolve the URIs that identify vocabulary terms in order to find their RDFS or OWL definition.
- The Web of Data is open
 - meaning that applications can discover new data sources at run-time by following links.

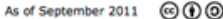
Linked Data Deployment on the Web



LoD Cloud in 2007



As of May 2007

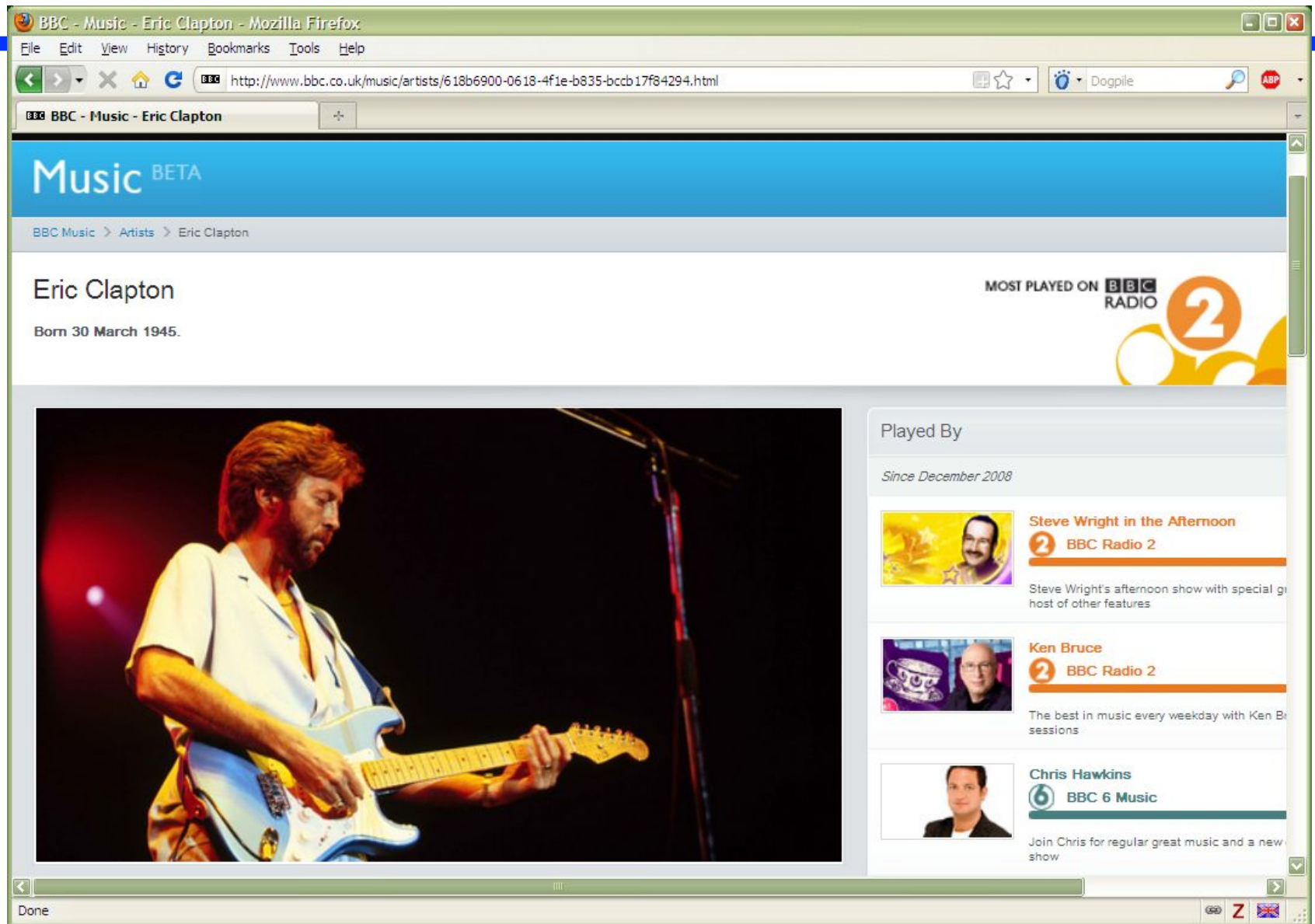


Some statistics from 2009

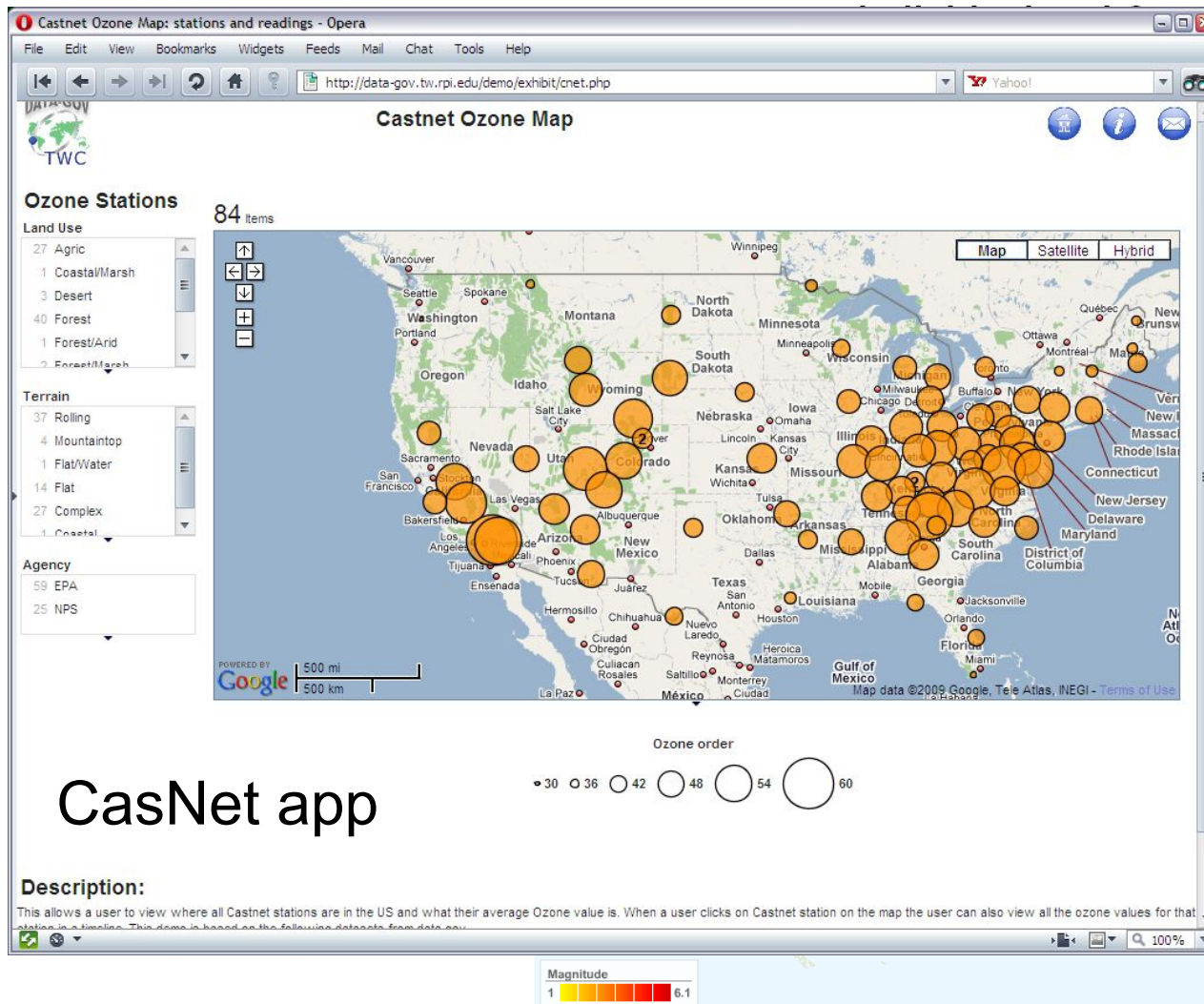
<i>Domain</i>	<i>No of Triples</i>	<i>% of Cloud</i>	<i>No of Links</i>	<i>% of Links</i>
Media	698.000.000	10,4%	1.238.000	0,8%
Publications	212.000.000	3,2%	4.922.000	3,3%
Life Sciences	2.429.000.000	36,1%	133.199.000	89,4%
Geographic Data	3.097.000.000	46,0%	4.038.000	2,7%
User Generate Content	76.000.000	1,1%	1.559.000	1,0%
Cross-Domain	214.000.000	3,2%	3.992.000	2,7%
<i>Total</i>	<i>6.726.000.000</i>		<i>148.948.000</i>	

+ 2 billion triples from Data.gov

Using the LOD to build Web sites: BBC

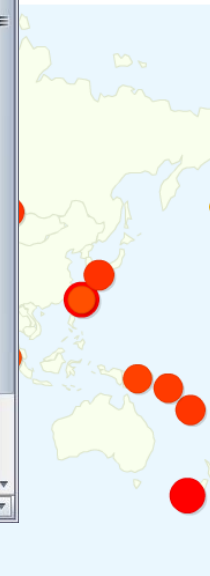
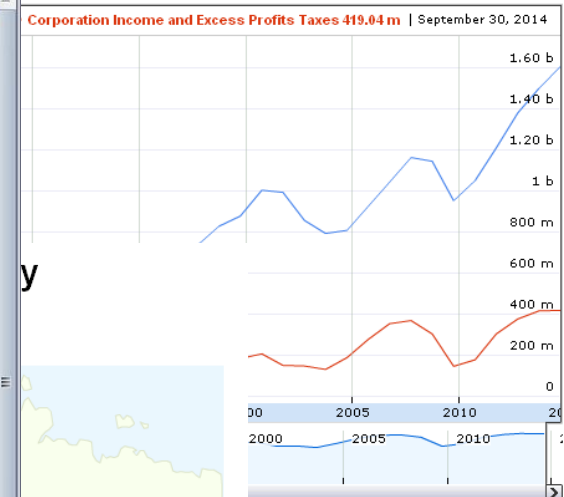


You publish raw data, others use it...



Corporate Tax Receipts

set 403



CastNet App Coding (1)

- Based on **2 linked data sets**
 - Clean Air Status and Trends Network (CASTNET): Ozone (from data.gov)
 - Dataset 10001: Castnet site information (from epa.gov)
- ... and **2 SPARQL queries**
 - Here are parts of the 1st one:

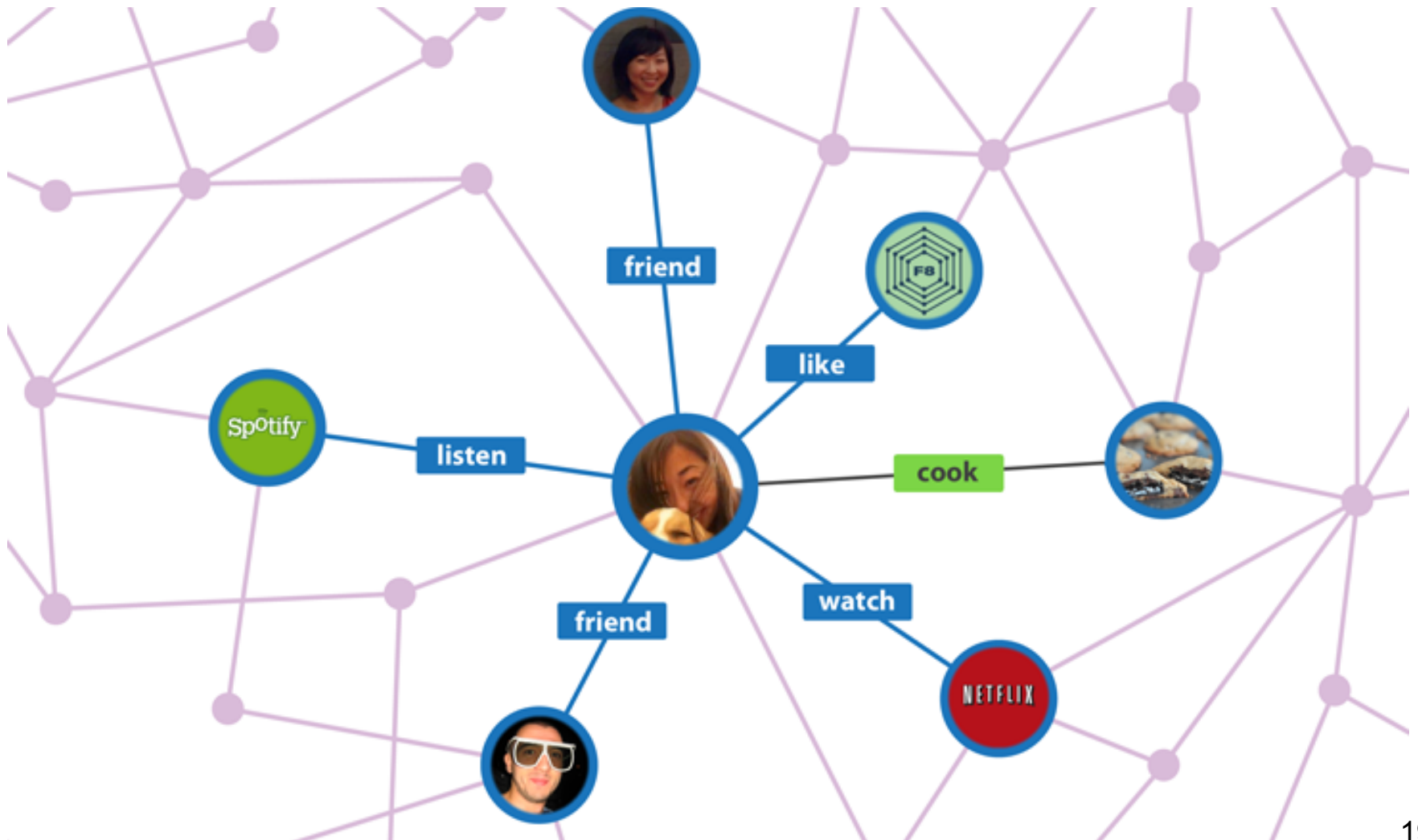
```
PREFIX dgp10001: <http://data-gov.tw.rpi.edu/vocab/p/10001/>
PREFIX dgp8: <http://data-gov.tw.rpi.edu/vocab/p/8/>
SELECT ... AVG ...
AVG(xsd:int(?ozone_8hr_daily_max)) AS ?average
  GRAPH <http://data-gov.tw.rpi.edu/vocab/Dataset_10001>
  {
    ?s dgp10001:latitude           ?lat .
    ?s dgp10001:longitude          ?long .
    ?s dgp10001:agency             ?agency .
    ?s dgp10001:site_name          ?label
  }
[...]
```

```
  GRAPH <http://data-gov.tw.rpi.edu/vocab/Dataset_8>
  {
    ?s2 dgp8:site_id              ?id .
    ?s2 dgp8:ozone_8hr_daily_max  ?ozone_8hr_daily_max
  }
}
```

CastNet App Coding (2)

- 20' development time
 - VS 2 months without Linked open Data
 - Identify the right data sets
 - Retrieve/Ask for the data sets
 - Understand/Parse formats
 - Manually integrate formats
 - Manually integrate data
 - Create links / joins
 - Store everything in a database
 - Build a dedicated javascript engine
 - Build a dedicated visualization package
 - And a dedicated GUI...
- ⇒ Automatization through crowdsourced, standardized data structures
- Can be manipulated by computational agents directly!

Facebook Open Graph

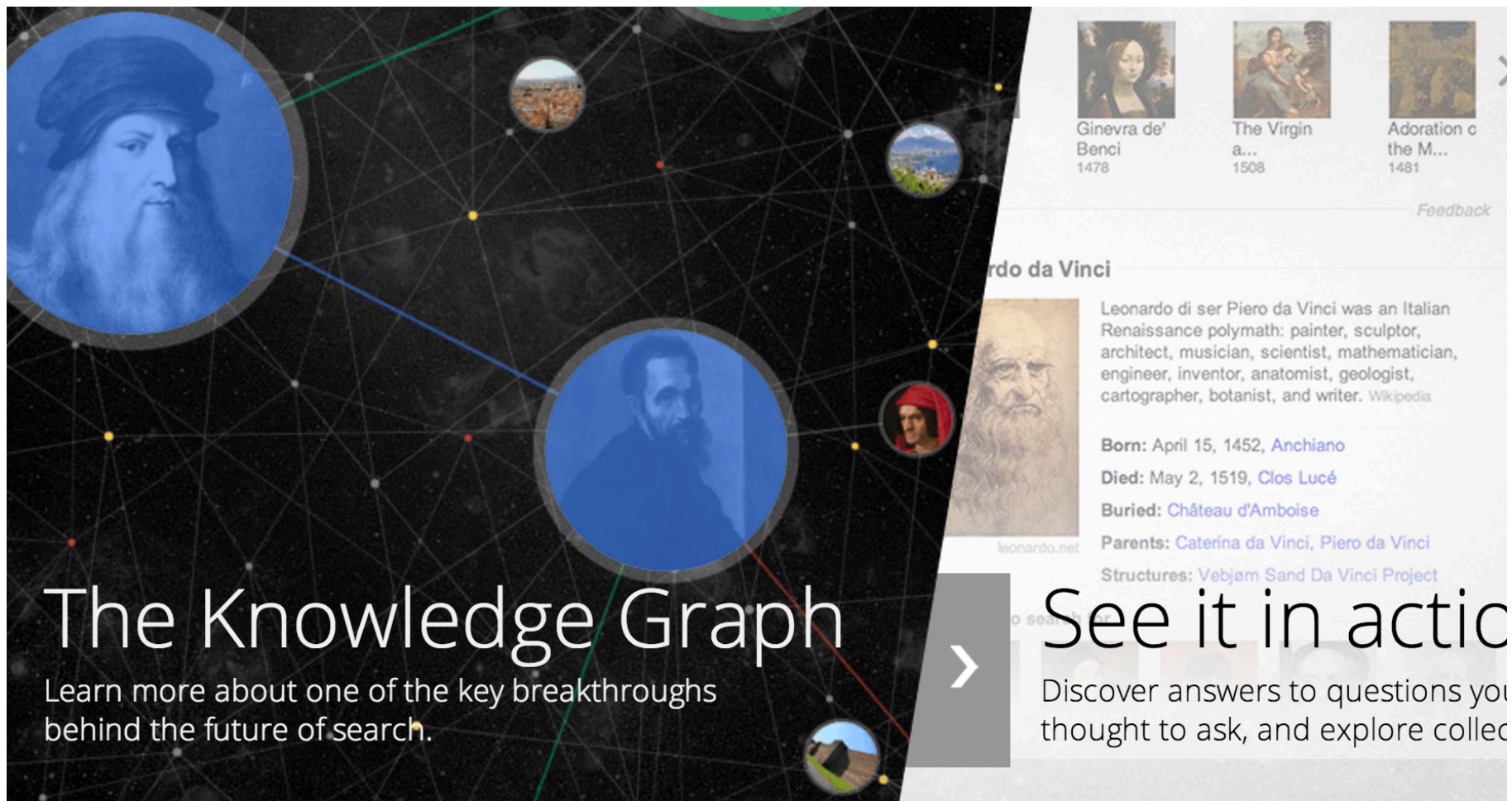


What Facebook Says about it

- “At Facebook's core is the social graph; people and the connections they have to everything they care about. Historically, Facebook has managed this graph and has expanded it over time as we launch new products (photos, places, etc.). In 2010, we extended the social graph, via the Open Graph protocol, to include 3rd party web sites and pages that people liked throughout the web. We are now extending the Open Graph to include arbitrary actions and objects created by 3rd party apps and enabling these apps to integrate deeply into the Facebook experience.”

Different Perspectives on LoD (1)

- Google's Knowledge Graph



The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

Leonardo da Vinci

Leonardo di ser Piero da Vinci was an Italian Renaissance polymath: painter, sculptor, architect, musician, scientist, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. Wikipedia

Born: April 15, 1452, [Anchiano](#)
Died: May 2, 1519, [Clos Lucé](#)
Buried: [Château d'Amboise](#)
Parents: [Caterina da Vinci](#), [Piero da Vinci](#)
Structures: [Vebjørn Sand Da Vinci Project](#)

See it in action

Discover answers to questions you thought to ask, and explore collections

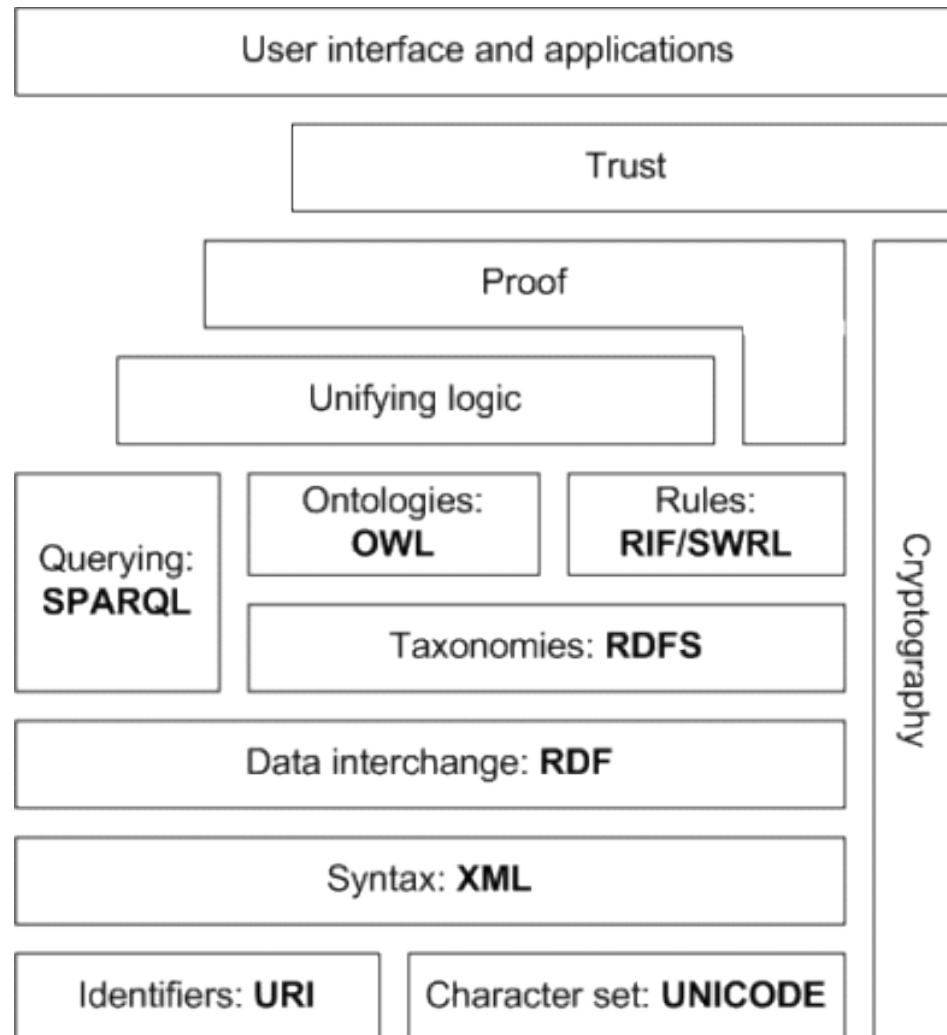
Different Perspectives on LoD (2)

- IBM Watson



- <http://www.youtube.com/watch?v=FC3IryWr4c8>
- <http://www.youtube.com/watch?v=DywO4zksfXw>

The Semantic Web Layer Stack



Resource Description Framework

- RDF, building block of the Semantic Web
- Used to encode data as *triples*, forming *distributed graphs*
- Standardized by the W3C
 - <http://www.w3.org/RDF/>
 - About a dozen recommendations (i.e., standards)
http://www.w3.org/standards/techs/rdf#w3c_all

RDF Triples

- Used to encode data in a semi-structured way
- Like small sentences: subject + attribute-value

1:subject, 2:predicate, 3:object

ex.: philippe made idmesh_paper:

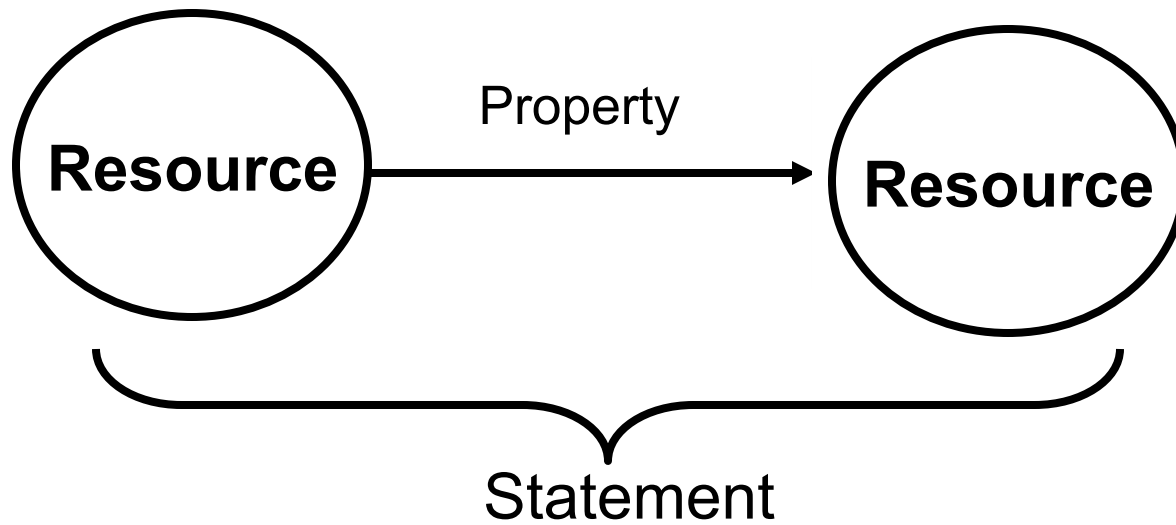
1: <http://data.semanticweb.org/person/philippe-cudre-mauroux>

2: <http://xmlns.com/foaf/0.1/made>

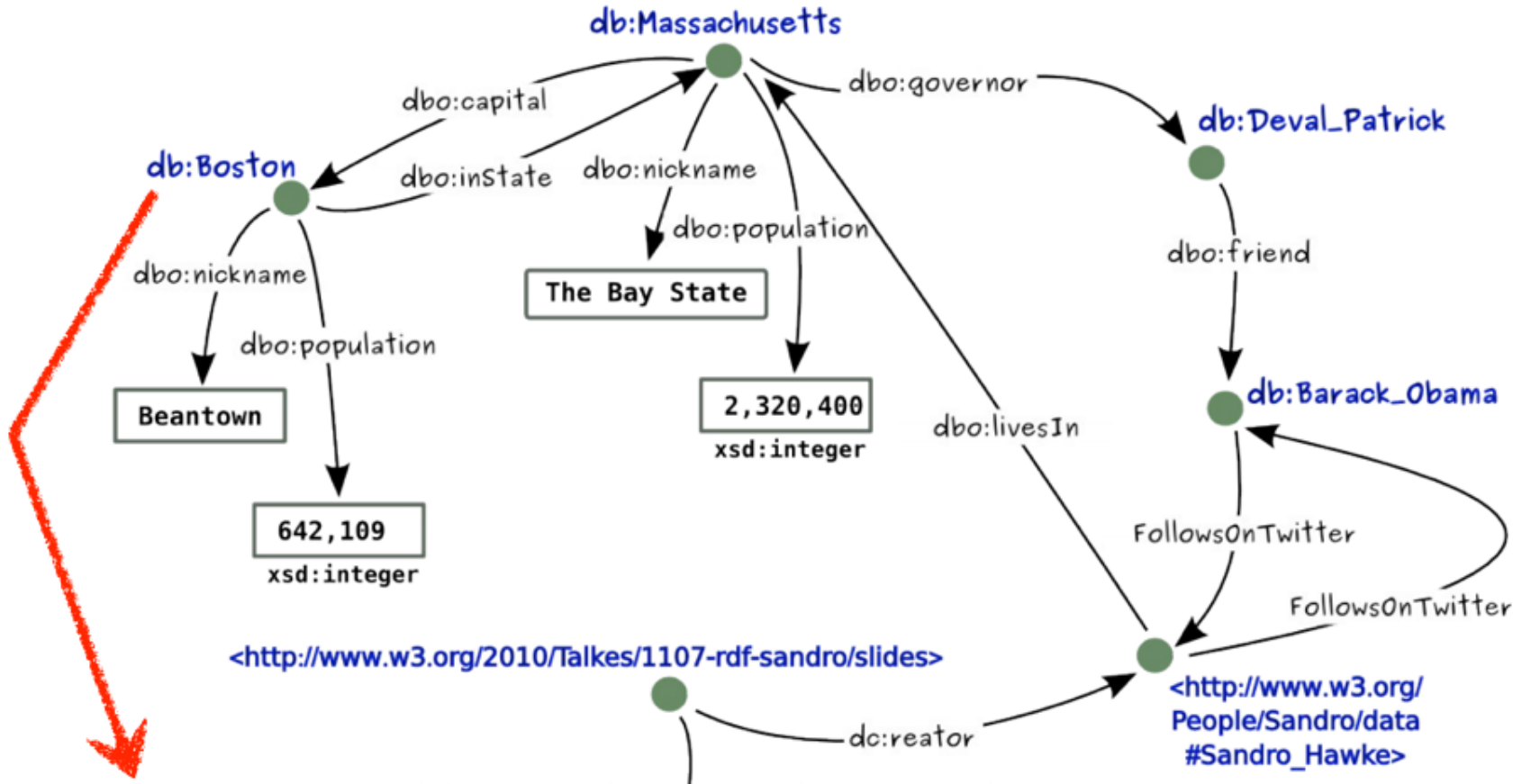
3: <http://data.semanticweb.org/conference/www/2009/paper/60>

- Subject: URI
- Predicate: URI
- Object: URI / value (“literal”)

RDF Model



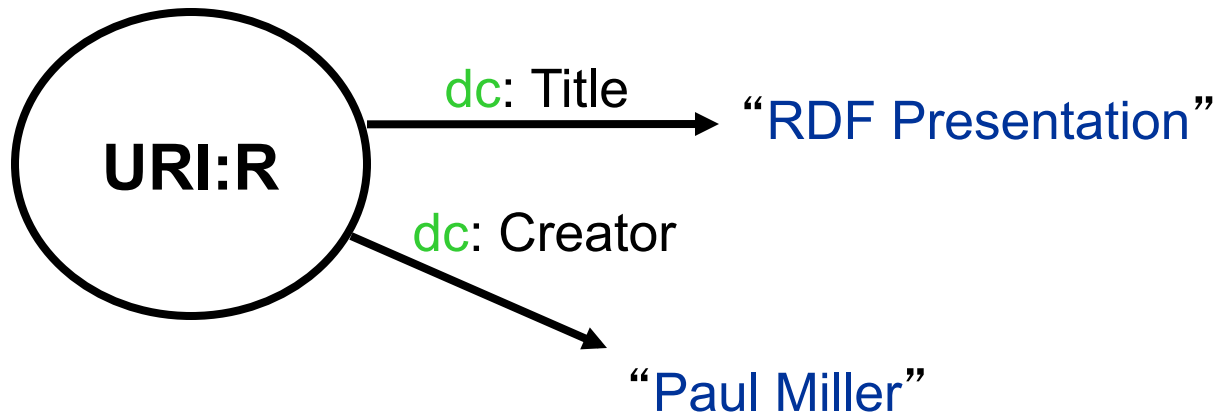
Naturally Forms Distributed Graphs



```
db:Boston dbo:nickname "Beantown".
db:Boston dbo:population "642109"^^xsd:integer.
db:Boston dbo:inState db:Massachusetts.
db:Massachusetts dbo:capital db:Boston.
db:Massachusetts dbo:nickname "The Bay State".
```

Graphs © Sandro Hawke, W3C

XML Serialization Example



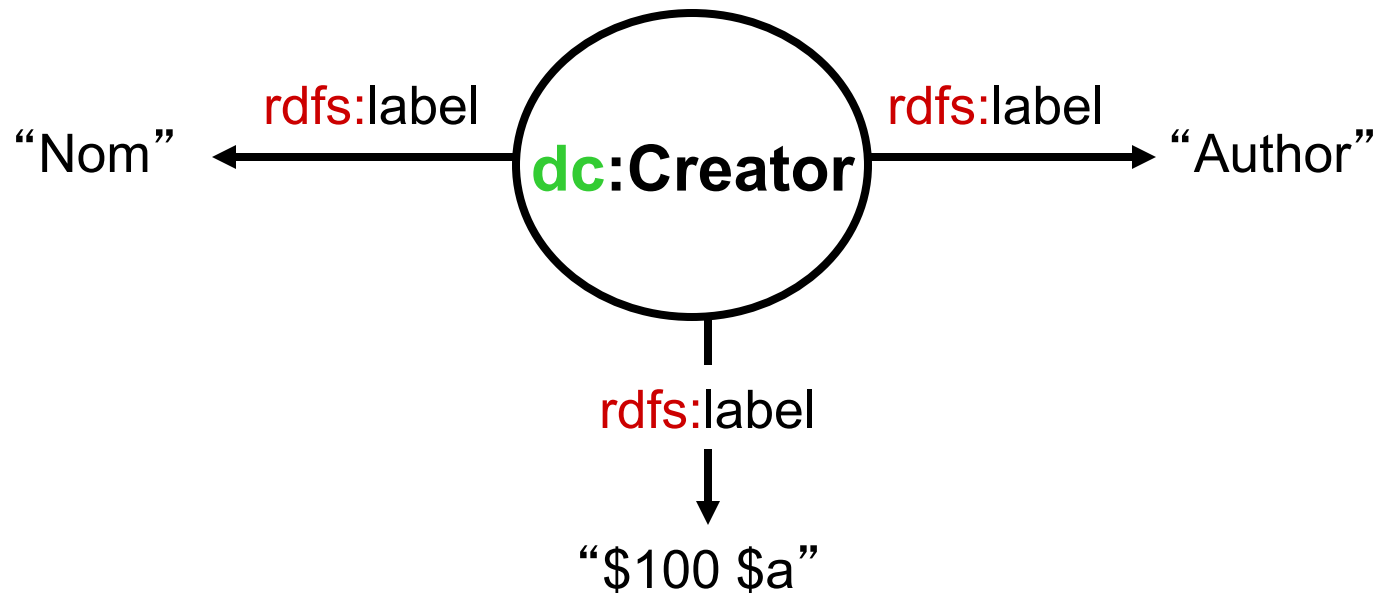
```
<RDF xmlns = "http://www.w3.org/TR/WD-rdf-syntax#"
      xmlns:dc = "http://purl.org/dc/elements/1.0/">
  <Description about = "URI:R">
    <dc:Title> RDF Presentation </dc:Title>
    <dc:Creator> Paul Miller </dc:Creator>
  </Description>
</RDF>
```

RDF Schemas

- Declaration of vocabularies
 - Properties/Classes defined by a particular community
 - characteristics of properties and/or constraints on corresponding values
- Expressible in the RDF model and syntax
- Provides structure!
 - easier to store, check, and process data

Schema Vocabularies

- Enables communities to share machine readable tokens and locally define human readable labels.



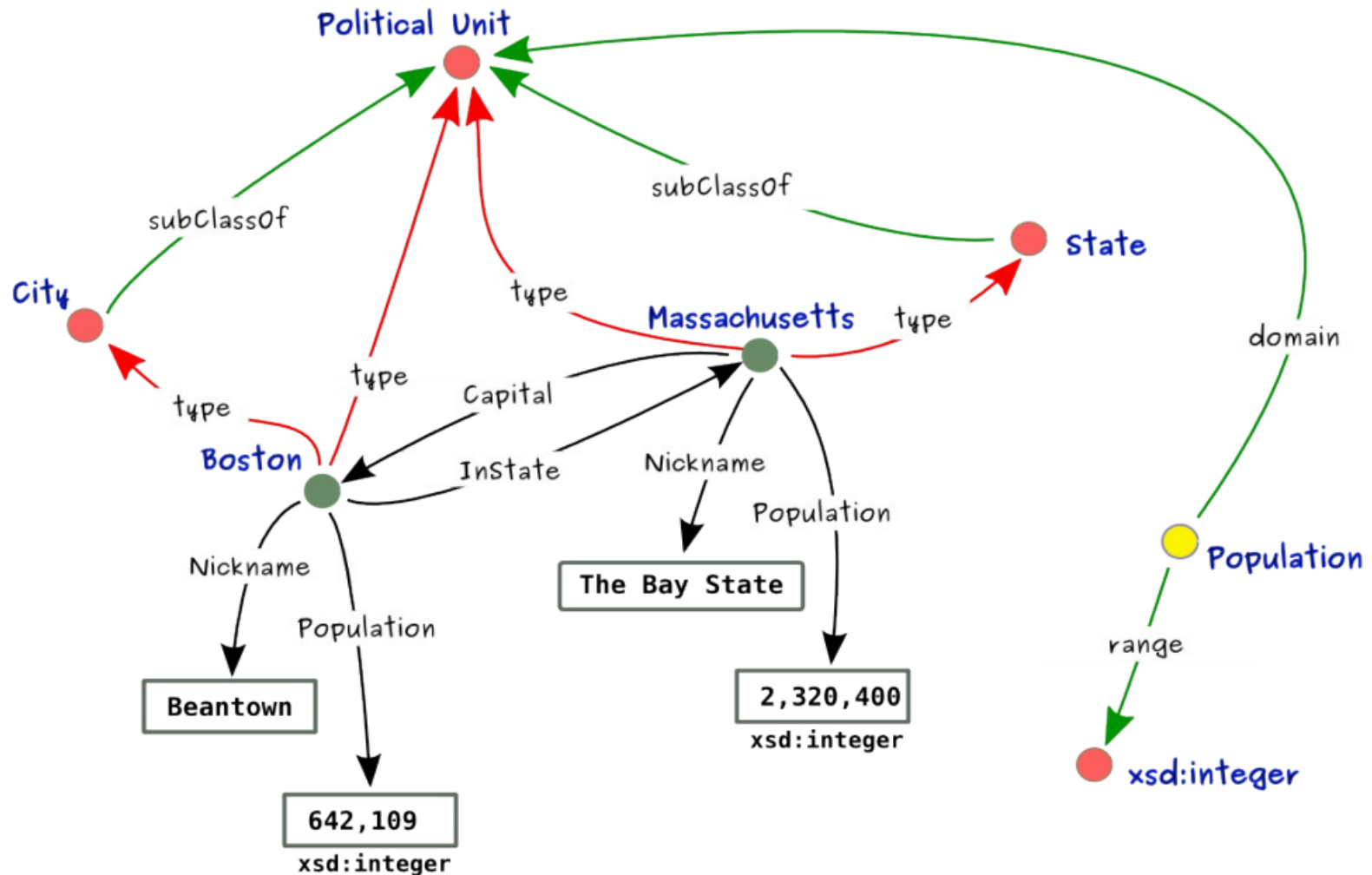
Examples of Vocabularies

- Friend of a Friend (Social Networks)
 - foaf:name
- Dublin Core (Publications)
 - dc:creator, dcterms:temporal
- Good Relations (Products)
 - gr:ProduceOrServiceModel, ...

RDF Schema Constructs

- Create classes
 - **Class** (as in OO, typed instances appearing as subjects/objects)
 - **Property** (typed predicates)
- Create hierarchies
 - **SubClassOf, SubPropertyOf**
 - Inheritance mechanisms
- Create constraints on the triples
 - **Domain** (restricts the subject of a property)
 - **Range** (restricts the object of a property)

RDF Schema Example



Live Example

- Draw an RDF schema + instance graph modeling students, teachers and courses

RDFa

- Embedding RDF information in HTML pages
Supported by Google, Yahoo, etc.

```
<body>
  <div about="http://dbpedia.org/resource/Massachusetts">The
    Massachusetts governor is
      <span rel="db:Governor">
        <span about="http://dbpedia.org/resource/Deval\_Patrick">Deval
          Patrick
        </span>,
      </span>
    the nickname is "<span property="db:Nickname">Bay State</span>",
    and the capital
    <span rel="db:Capital">
      <span about="http://dbpedia.org/resource/Boston">
        has the nickname "<span property="db:Nickname">Beantown</span>".
      </span>
    </span>
  </div>
</body>
```

OWL

- The Web Ontology Language
 - Very expressive schemas! (ontologies)
 - Description Logics
 - ... and several flavours
 - Example: OWL 2 EL axioms:
 - class inclusion (SubClassOf)
 - class equivalence (EquivalentClasses)
 - class disjointness (DisjointClasses)
 - object property inclusion (SubObjectPropertyOf) with or without property chains, and data property inclusion (SubDataPropertyOf)
 - property equivalence (EquivalentObjectProperties and EquivalentDataProperties),
 - transitive object properties (TransitiveObjectProperty)
 - reflexive object properties (ReflexiveObjectProperty)
 - domain restrictions (ObjectPropertyDomain and DataPropertyDomain)
 - range restrictions (ObjectPropertyRange and DataPropertyRange)
 - assertions (SameIndividual, DifferentIndividuals, ClassAssertion, ObjectPropertyAssertion, DataPropertyAssertion, NegativeObjectPropertyAssertion, and NegativeDataPropertyAssertion)
 - functional data properties (FunctionalDataProperty)
 - keys (HasKey)
 - Inference! ex.: TransitiveObjectProperty(hasAncestor)
 $\text{hasAncestor}(x, y) \wedge \text{hasAncestor}(y, z) \rightarrow \text{hasAncestor}(x, z)$

SPARQL

- Declarative query language for RDF/S
 - SQL for the Semantic Web!

```
prefix db: <http://dbpedia.org/resource/>
prefix dbo: <http://dbpedia.org/ontology/>

SELECT ?cap
WHERE { db:Massachusetts dbo:capital ?cap }
```

- Uses *triple patterns*
 - *?subject ?predicate ?object*

More Complex SPARQL Query

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title2
WHERE
{
    ?doc      dc:title      "SPARQL at speed" .
    ?doc      dc:creator    ?c .
    ?docOther dc:creator    ?c .
    ?docOther dc:title      ?title2
}
```

- On an abstracts/papers database:
“Find other papers by the authors of a given paper.”

Optional graph patterns

Data

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .  
@prefix : <http://example.org/book/> .  
@prefix ns: <http://example.org/ns#> .  
:book1 dc:title "SPARQL Tutorial" .  
:book1 ns:price 42 .  
:book2 dc:title "The Semantic Web" .  
:book2 ns:price 23 .
```

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>  
PREFIX ns: <http://example.org/ns#>  
SELECT ?title ?price  
WHERE { ?x dc:title ?title .  
        OPTIONAL { ?x ns:price ?price .  
                  FILTER ?price < 30 }}
```

Query

Query Result

title	price
"SPARQL Tutorial"	
"The Semantic Web"	23

SPARQL Constructs

- Many other constructs
 - Order By
 - Distinct
 - Limit
 - Construct
 - Ask
 - Value tests
 - Transitive closures!

References

- <http://linkeddata.org/>
 - <http://developers.facebook.com/docs/opengraph/>
 - <http://data-gov.tw.rpi.edu/>
 - <http://www.w3.org/2001/sw/>
-
- Many of the slides in this deck were adapted from presentations by Ivan Herman (W3C) and Chris Bizer (Freie U. Berlin)