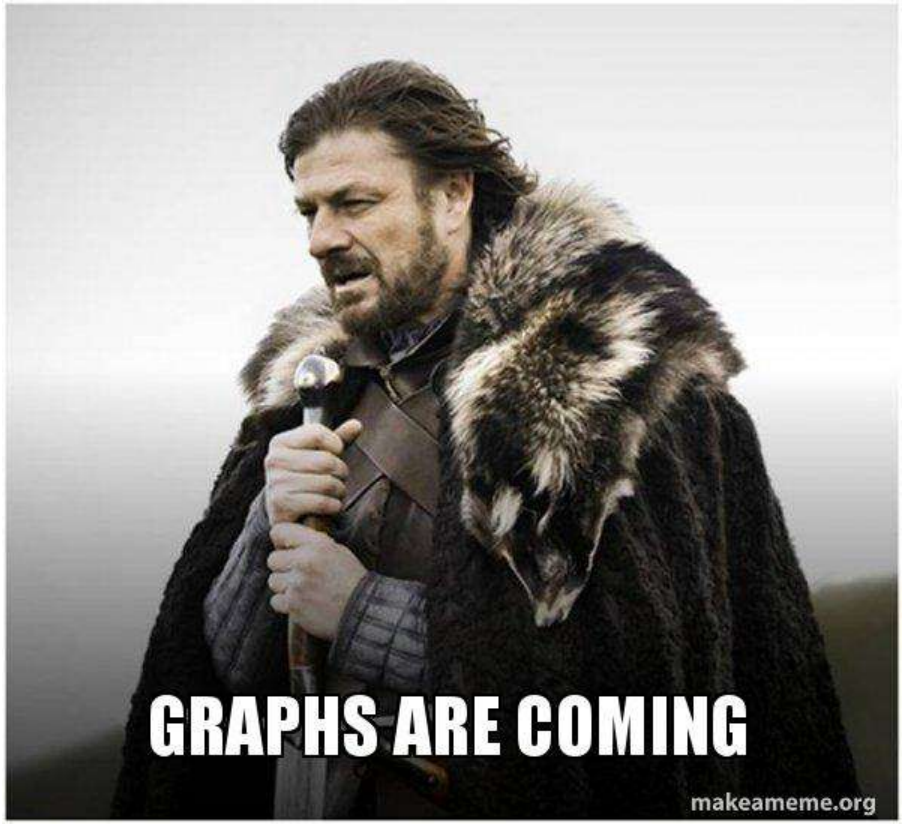
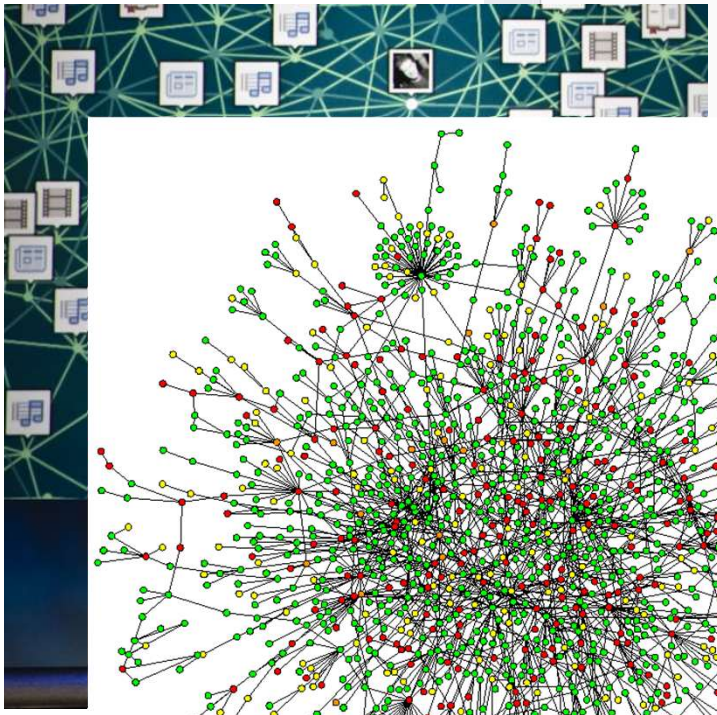
Bases de datos de grafos

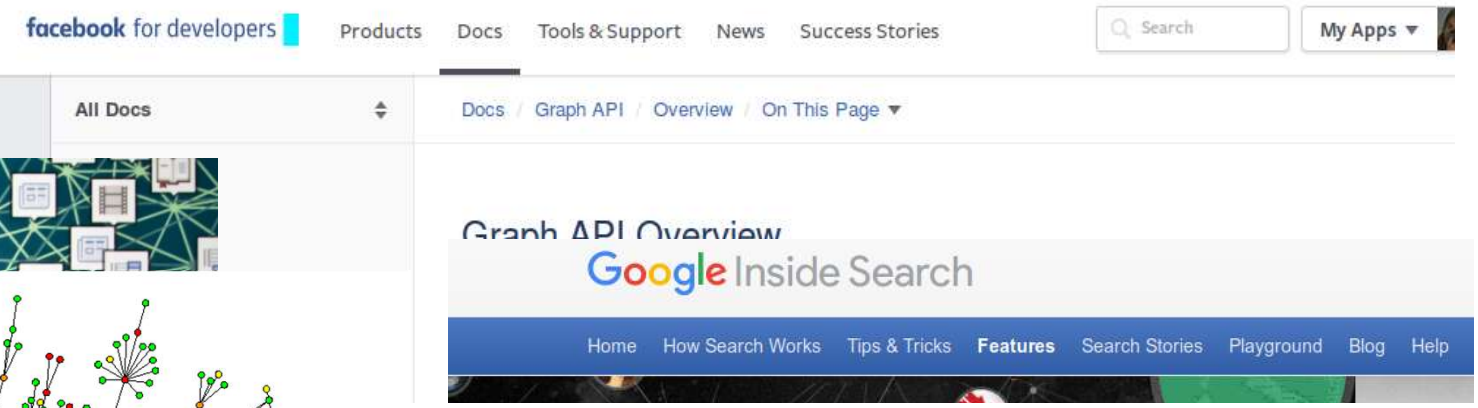


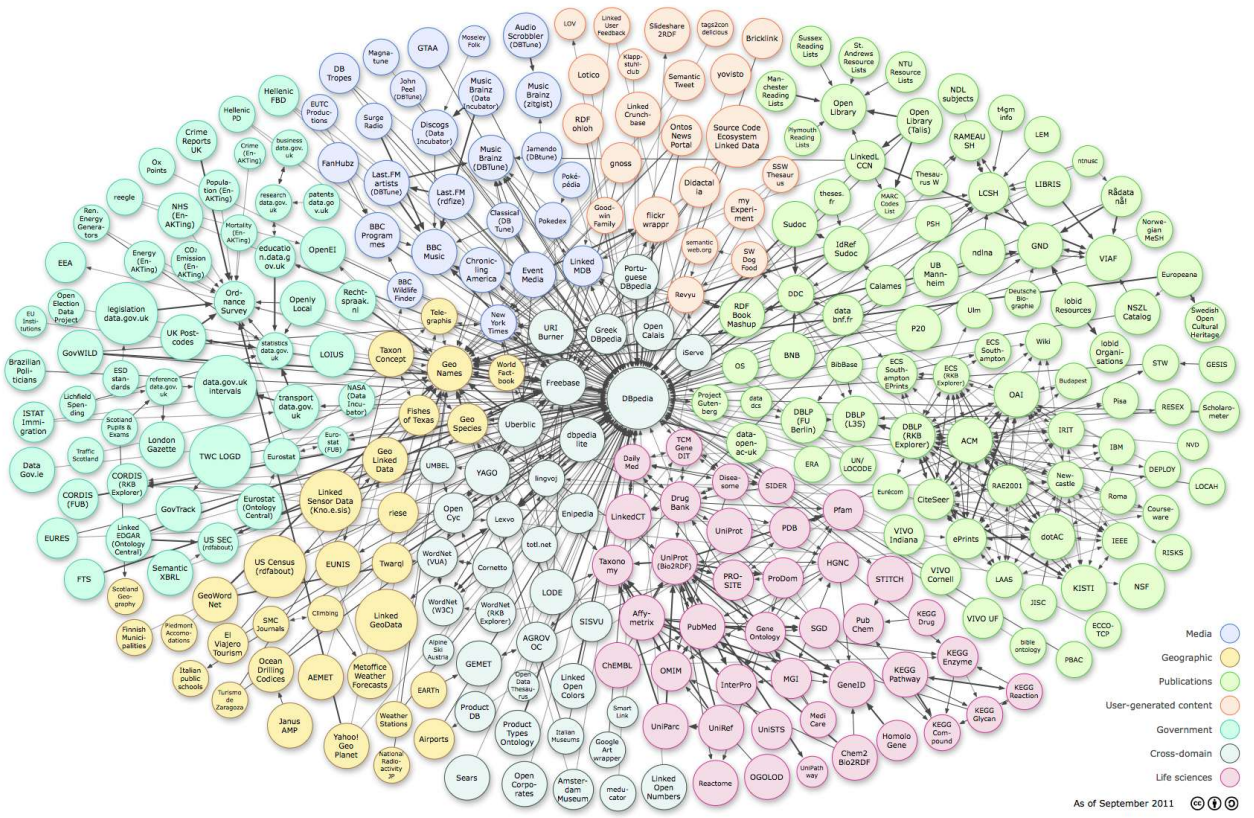
Instituto de Computación, FING, UdelaR

CC-BY Lorena Etcheverry lorenae@fing.edu.uy

Hay casos en que interesa más modelar las relaciones entre cosas que las cosas!

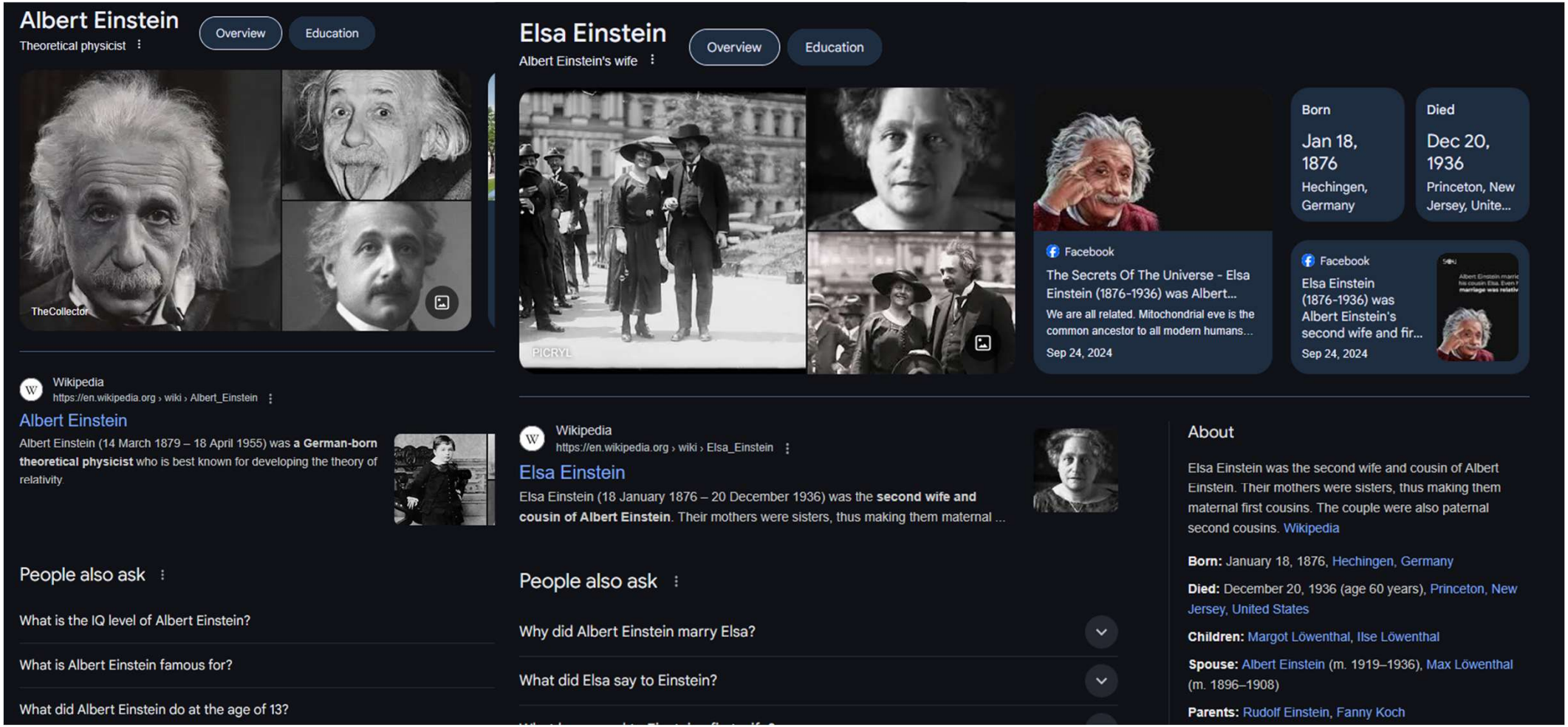






Knowledge Graphs

"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." (Hogan et al, Knowledge Graphs, 2021)



https://support.google.com/knowledgepanel/answer/9787176?hl=en

¿en qué escenarios son adecuadas las BD de grafos?

- Las entidades están muy conectadas a través de relaciones descriptivas.
- Existen relaciones cíclicas o entidades autorreferenciadas.
  - Esto suele ser un reto cuando se utilizan bases de datos relacionales o documentales.
- Las relaciones entre entidades evolucionan dinámicamente.
  - especialmente aplicable a datos jerárquicos o estructurados en árbol con muchos niveles.
- Existen relaciones de muchos a muchos entre las entidades.
- Existen requisitos de escritura y lectura tanto en las entidades como en las relaciones.

Agenda

- Modelos en bases de datos de grafos
  - El modelo *Property graph*
  - RDF y modelos en la web semántica
  - Diseño de bases de datos de grafos

- Motores de bases de datos de grafos
  - Estrategias de almacenamiento
  - Lenguajes de consulta
  - Modelos de procesamiento distribuído

Modelos en bases de datos de grafos

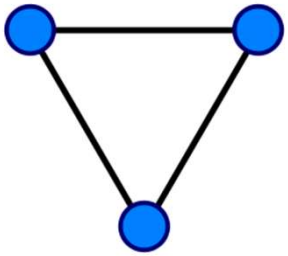
¿Qué es un grafo?

Un grafo G consiste en
un conjunto de nodos o vértices V,
y un conjunto de aristas E.
Las aristas conectan nodos entre si.

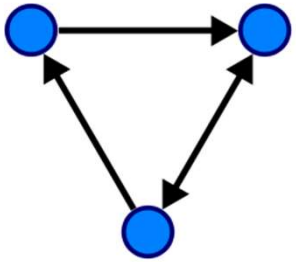
Las bases de datos de grafos
implementan diferentes variantes[1].

[1] *Survey of Graph Database Models*, Angles and Gutierrez,
ACM Computing Surveys, 2008

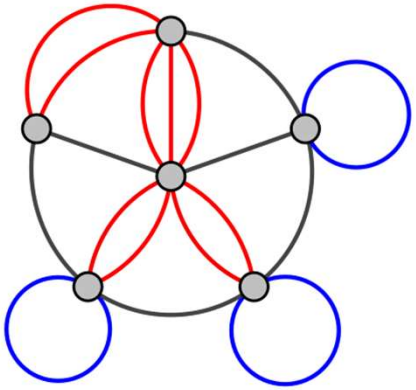
Grafos según su estructura



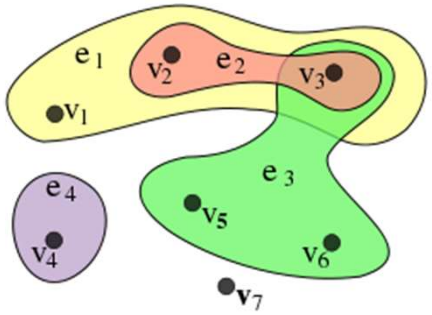
Grafo no dirigido



Grafo dirigido

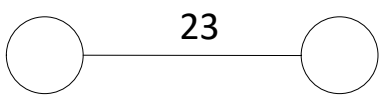


Multigrafos y pseudografos



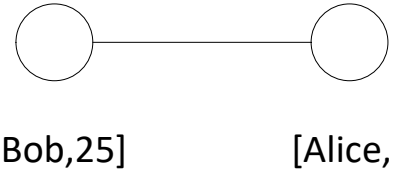
Hipergrafos

Grafos según los datos asociados



Etiquetas en las aristas (Edge-labelled)

Etiquetas en los nodos



Atributos en los nodos

¿qué tipos de grafos se utilizan en las bases de datos de grafos?

Existen muchos modelos (ver *Survey of Graph Database Models*, Angles and Gutierrez, ACM Computing Surveys, 2008)

Los sistemas más populares actualmente utilizan Property Graph [Miller, 2013] o Resource Description Framework [Cyganiak et al., 2014]

Grafo dirigido con etiquetas en las aristas

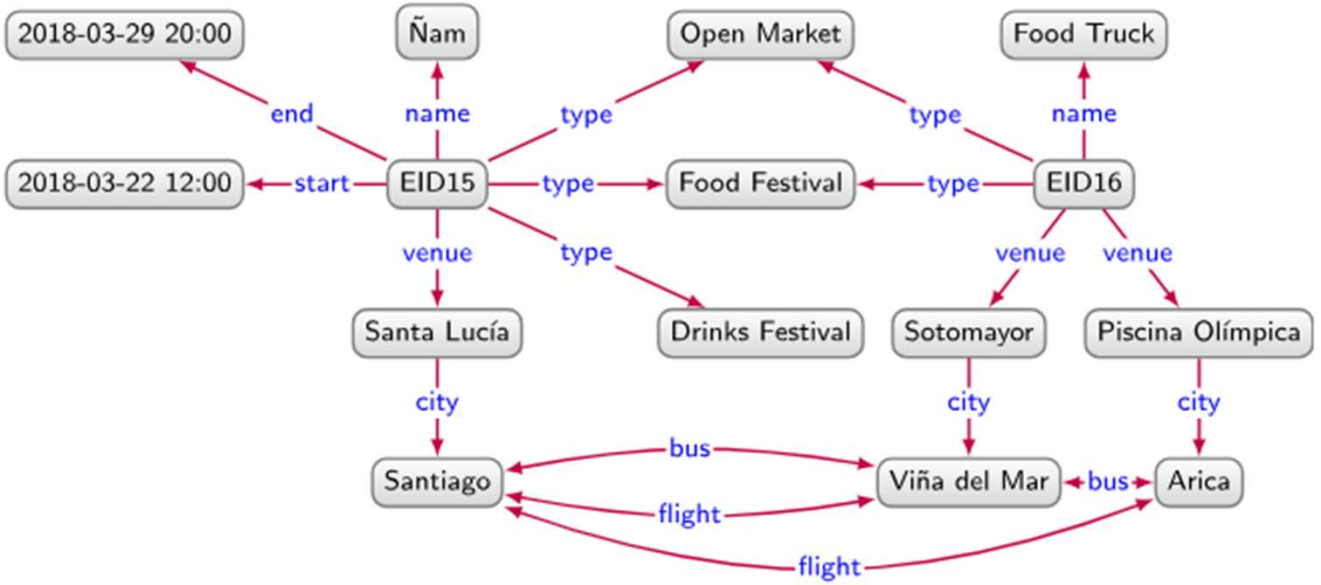


Figure 2.1: Directed edge-labelled graph describing events and their venues

El modelo RDF está basado en este tipo de grafos

Definition 2.1 (Directed edge-labelled graph)

A directed edge-labelled graph is a tuple $G = (V, E, L)$*, where* $V \subseteq$ **Con** *is a set of nodes,* $L \subseteq$ **Con** *is a set of edge labels, and* $E \subseteq V \times L \times V$ *is a set of edges.*

Resource Description Framework (RDF)

Pseudografos dirigidos.

Etiquetas en nodos y aristas.

Las etiquetas pueden ser:
IRIs, literales, o nodos blancos.

Cada nodo o arista tiene una sola etiqueta

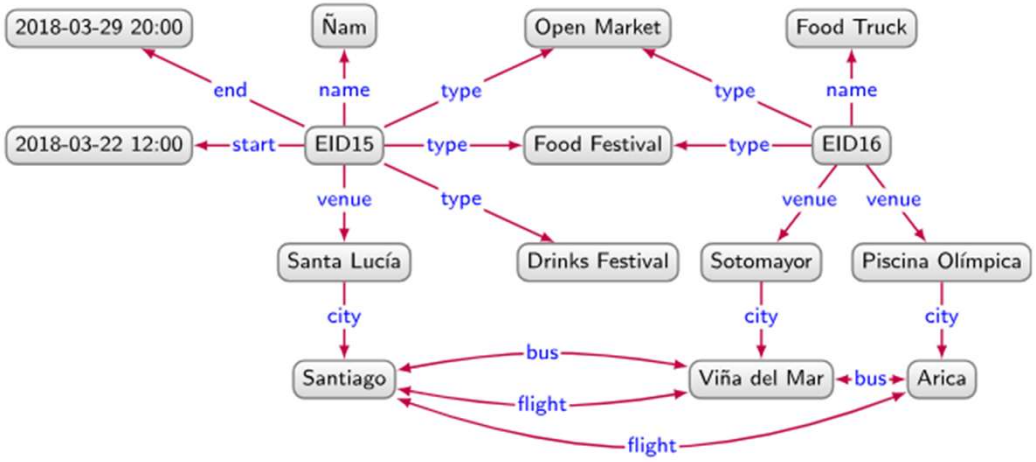
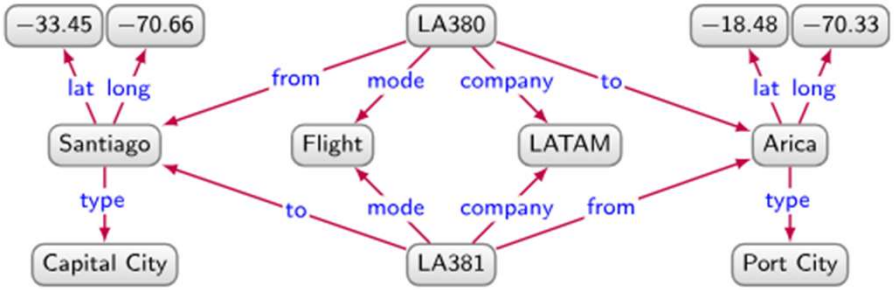


Figure 2.1: Directed edge-labelled graph describing events and their venues

Virtuoso y Stardog son ejemplos de sistemas de bases de datos que soportan RDF.

Property graph model



(a) Directed edge-labelled graph

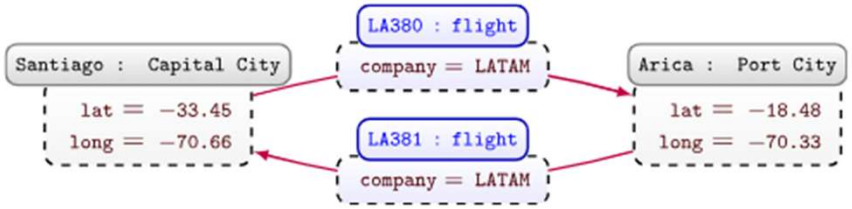


Figure 2.3: C

Definition 2.3 (Property graph)

A property graph is a tuple $G = (V, E, L, P, U, e, l, p)$*, where* $V \subseteq$ **Con** *is a set of node ids,* $E \subseteq$ **Con** *is a set of edge ids,* $L \subseteq$ **Con** *is a set of labels,* $P \subseteq$ **Con** *is a set of properties,* $U \subseteq$ **Con** *is a set of values,* $e : E \to V \times V$ *maps an edge id to a pair of node ids,* $l : V \cup E \to 2^L$ *maps a node or edge id to a set of labels, and* $p : V \cup E \to 2^{P \times U}$ *maps a node or edge id to a set of property–value pairs.*

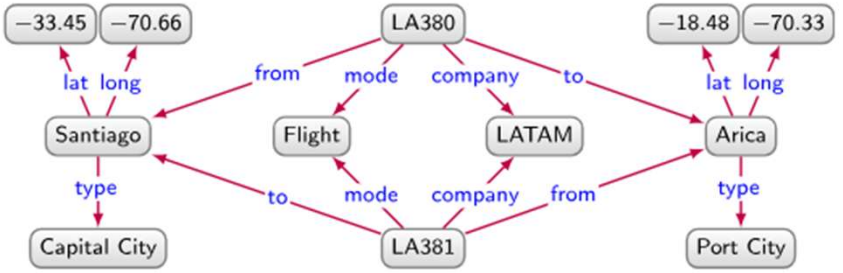
Property Graph Model (PGM)

Pseudografos dirigidos.

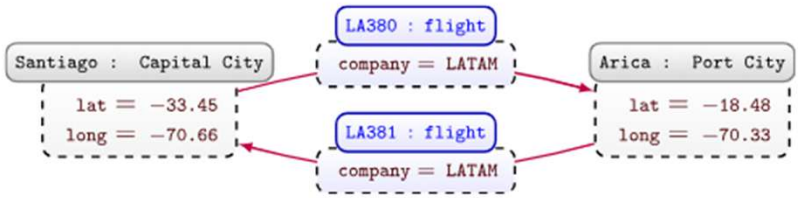
Parejas (clave,valor) llamadas *propiedades* asociadas a nodos y aristas.

Puedo **etiquetar** nodos y aristas.

Cada nodo o arista puede tener más de una etiqueta



(a) Directed edge-labelled graph

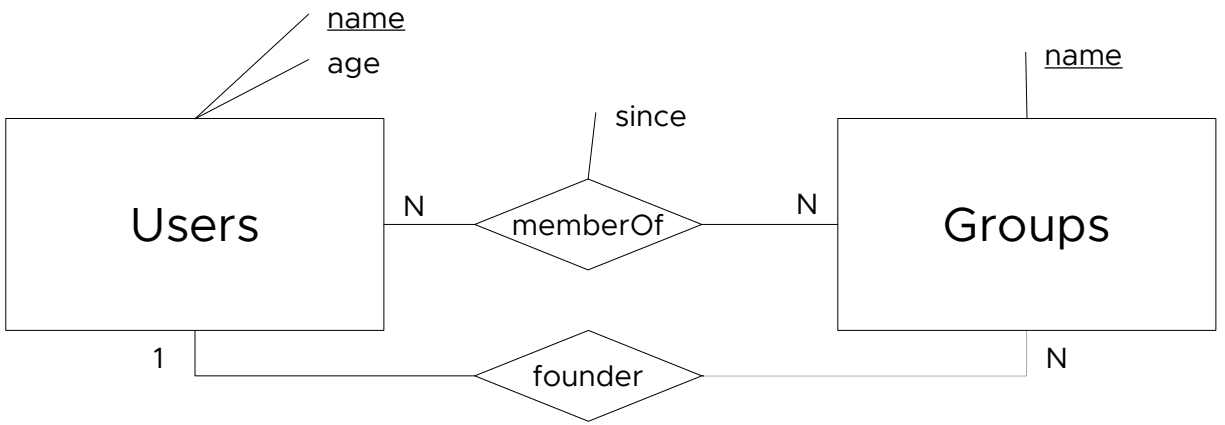


(b) Property graph

Figure 2.3: Comparing directed edge-labelled graphs and property graphs

Neo4j y Titan son ejemplos de sistemas de bases de datos que soportan PGM.

CONCEPTUAL



LÓGICO

Relacional

Users

name	age
Bob	25
Alice	null

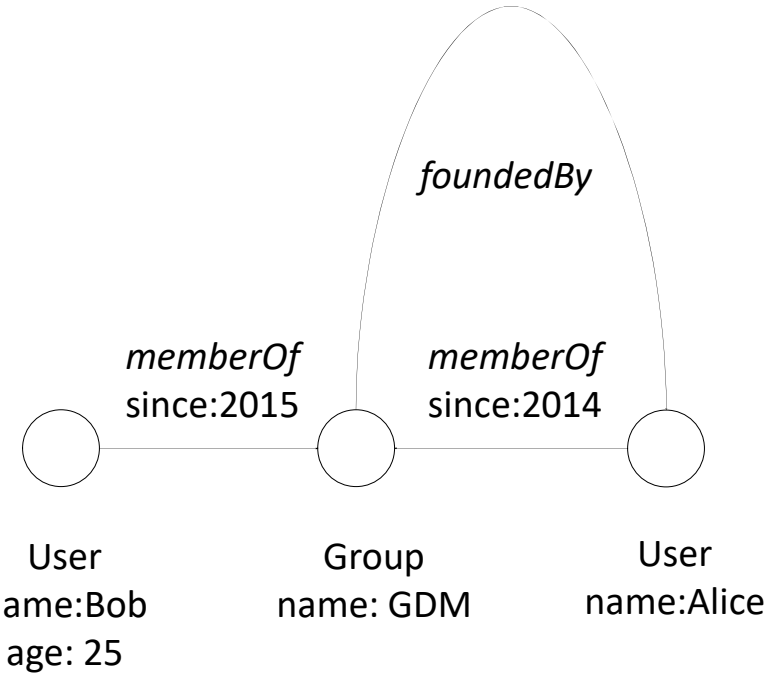
Groups

name	founder
GDM	Alice

GroupMembers

userName	groupName	since
Bob	GDM	2015
Alice	GDM	2014

Property Graphs



Restricciones y esquemas

- En ambos casos es posible definir nociones de esquema más o menos fuertes.
- En la familia de lenguajes y modelos vinculados a RDF puedo hacer cosas muy expresivas y potentes.
- Sobre el PGM puedo definir en forma teórica algunas restricciones pero dependen del vendedor
  - Ejemplo Neo4j constraints

Slide 16

1 Acá hay que agregar info, inclusive más slides
, 5/6/2023

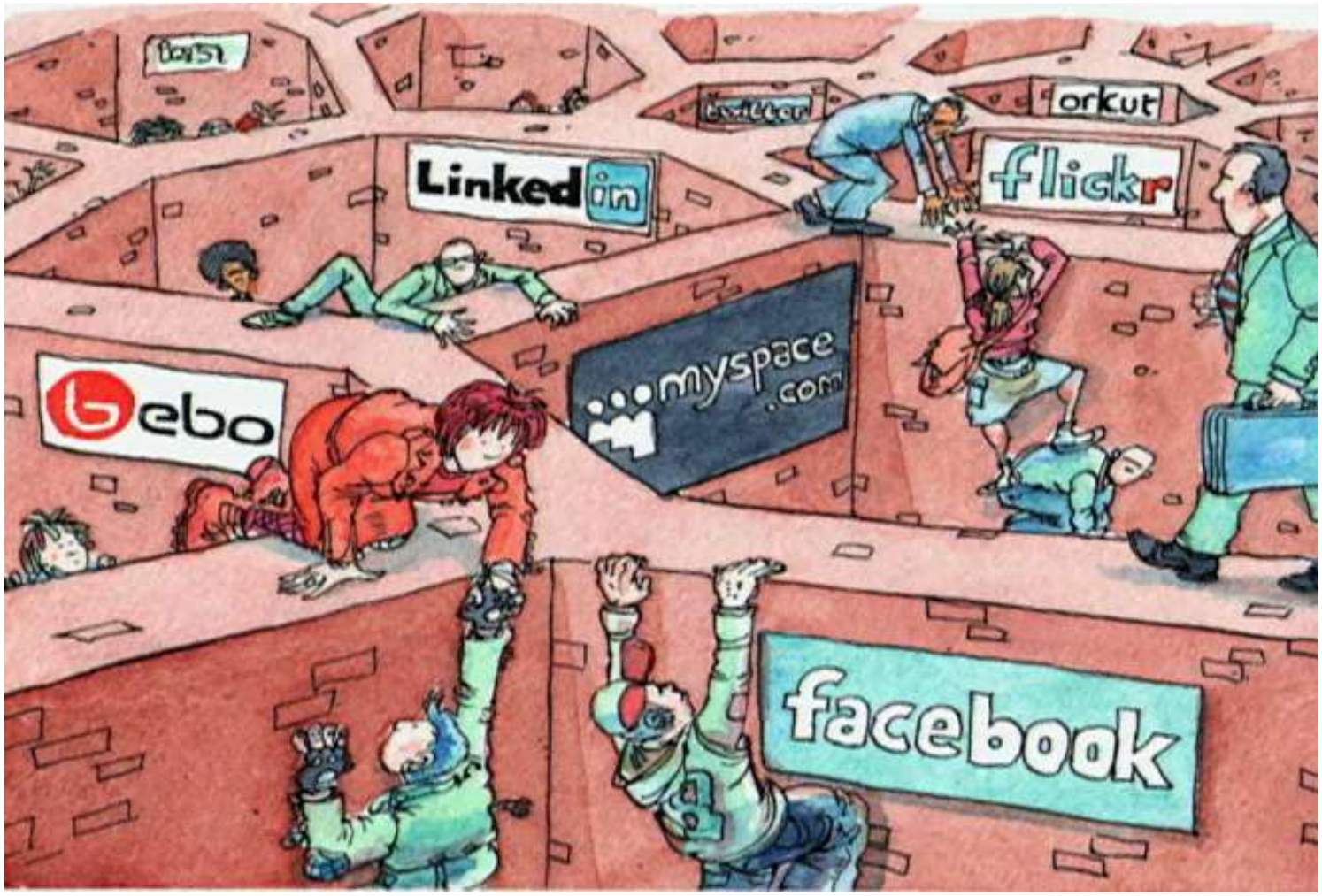
Grafos de conocimiento y la web semántica

La web clásica es una red de documentos,

interpretable por humanos,

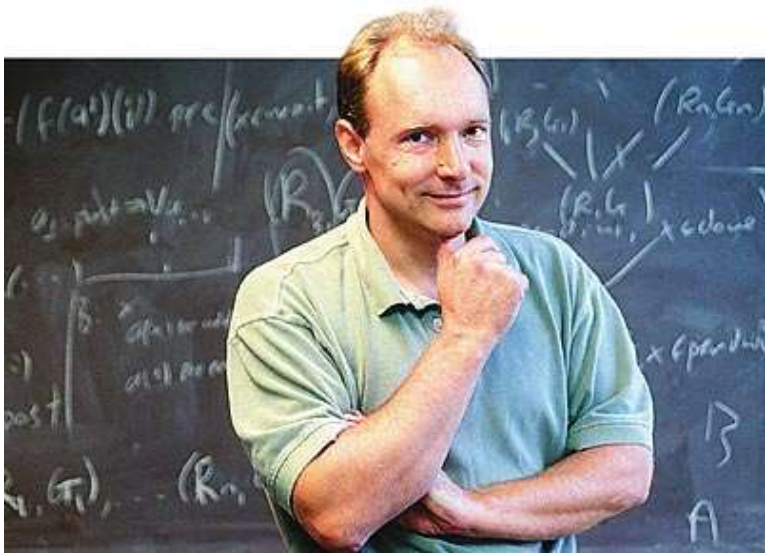
donde las relaciones entre documentos no tienen un significado.

Además, la mayoría de los datos en la web están a i s l a d o s



La web de datos es una red de afirmaciones, interpretable por humanos y máquinas, donde las afirmaciones se relacionan y tienen un significado.

W3C Semantic Web



RDF: un modelo de datos basado en triplas que permite representar relaciones.

SPARQL: el lenguaje de consultas sobre RDF.

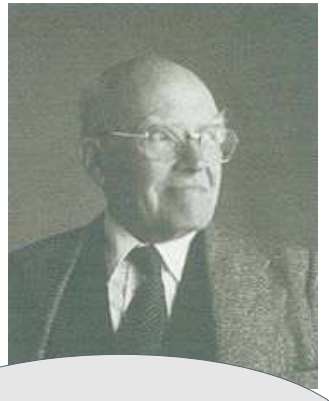
RDF-S y OWL, para representar metadatos y darles significado (ontologías).

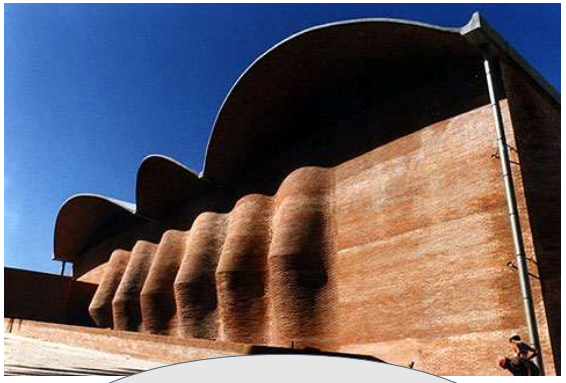
http://www.w3.org/2001/sw/wiki/RDF

http://www.w3.org/2001/sw/wiki/SPARQL

afirmación o tripla

<sujeto, predicado, objeto>





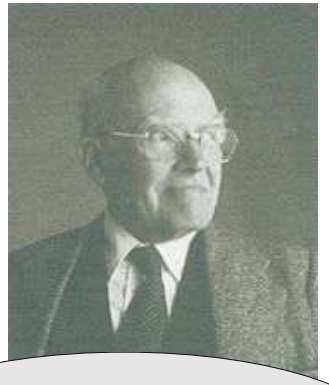
Ing. Eladio Dieste

esAutor

MVD Shopping Center

afirmación o tripla

<sujeto, predicado, objeto>



http://.../Eladio_Dieste — esAutor — http://.../Montevideo_Shopping — ubicadoEn — http://.../Buceo

En RDF todo es una tripla.
Los recursos y las propiedades se identifican por URIs

Linked Data

es un conjunto de
buenas prácticas para
publicar y *relacionar*
datos en la web,
usando
tecnologías de la
web semántica

Principios de Linked Data

1) Usar URIs para nombrar cosas

2) Usar URIs HTTP que sean consultables por humanos y máquinas

3) Proveer información útil acerca de cada URI en RDF

4) Crear links entre URIs

Tim Berners-Lee – Linked Data (2006)
http://www.w3.org/DesignIssues/LinkedData
#TED Talk Tim Berners-Lee on the next Web (2009)http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

1) Usar **URIs** para nombrar cosas

URI's *minting*: the act of establishing the association between the URI and the resource it denotes. A URI MUST only be minted by the URI's owner or delegate. Minting a URI from someone else's URI space is known as URI *squatting*

The URI Lifecycle in Semantic Web Architecture, David Booth, http://dbooth.org/2009/lifecycle/

1) Usar **URIs** para nombrar cosas

- Evitar el uso de *namespaces* que no están bajo nuestro control

  http://www.imdb.com/title/tt0057012/#film

- No incluir detalles de implementación

  http://tiger.biglynx.co.uk/people.php?id=dave-smith&format=rdf

- Evitar claves “artificiales”

  http://data.nytimes.com/66209802438676211043

2) Usar **URIs HTTP** que sean consultables por humanos y máquinas

HTTP URI: mecanismo universal de acceso a una URI

Las URIs HTTP deberían ser

desreferenciables

"HTTP clients can look up the URI using the HTTP protocol and retrieve a description of the resource that is identified by the URI"

Heath, Bizer, *Linked Data: Evolving the Web into a Global Data Space*

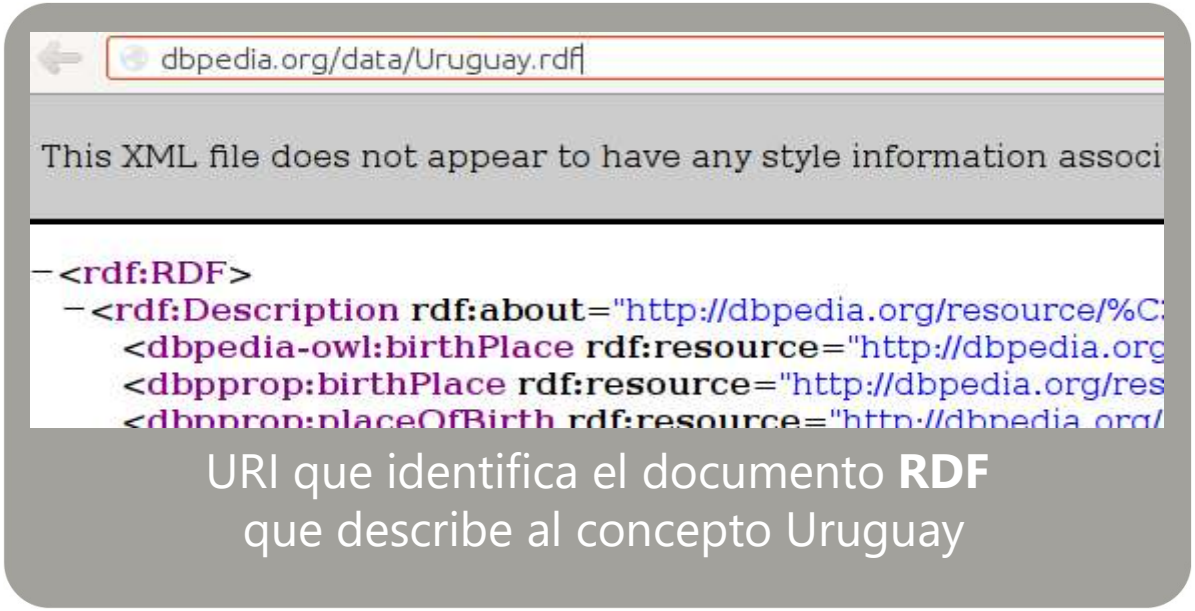
Redireccionamiento y negociación de contenido

text/html

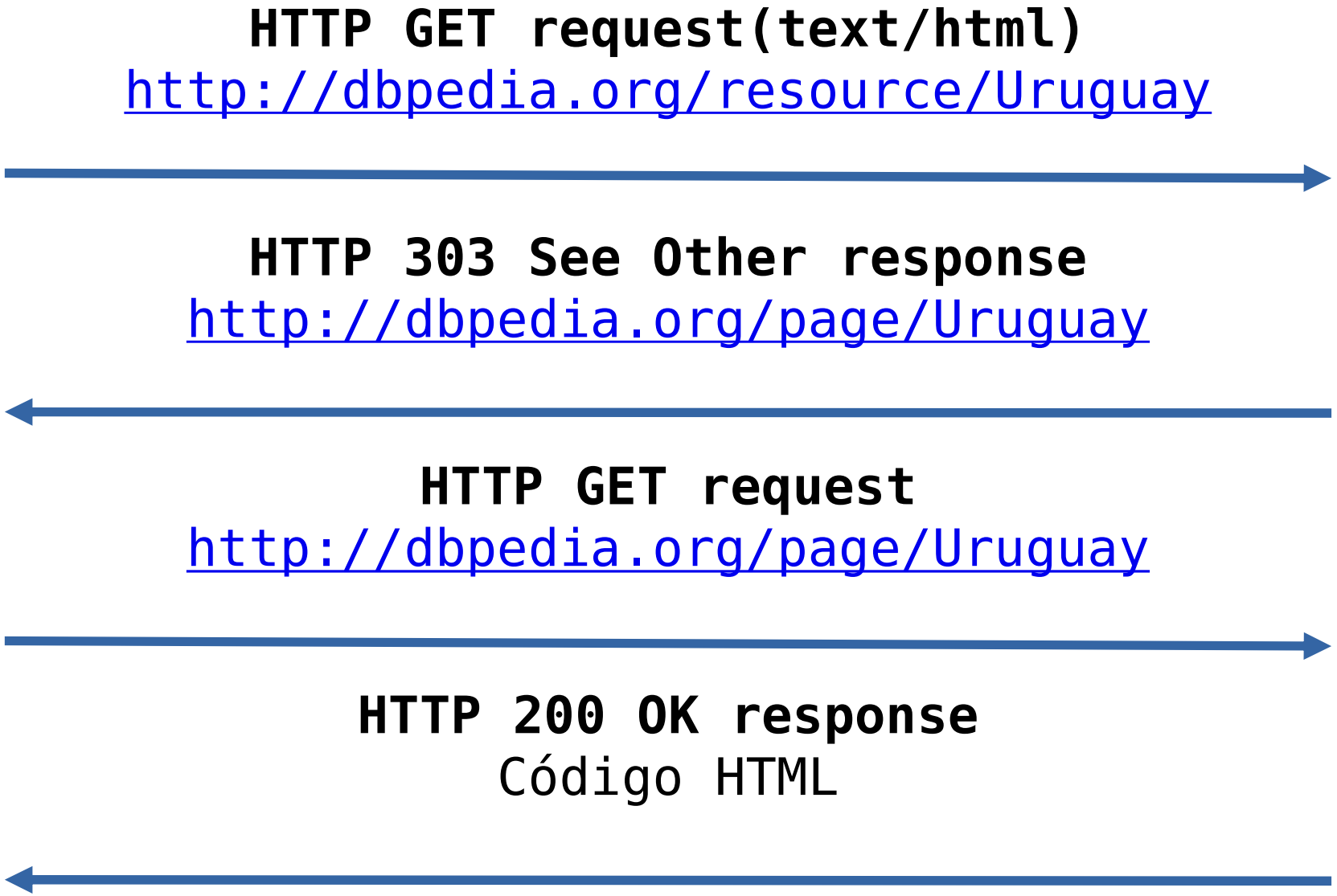
URI que identifica al concepto Uruguay

URI que identifica el documento **HTML** que describe al concepto Uruguay

application/rdf+xml



URI que identifica el documento **RDF** que describe al concepto Uruguay

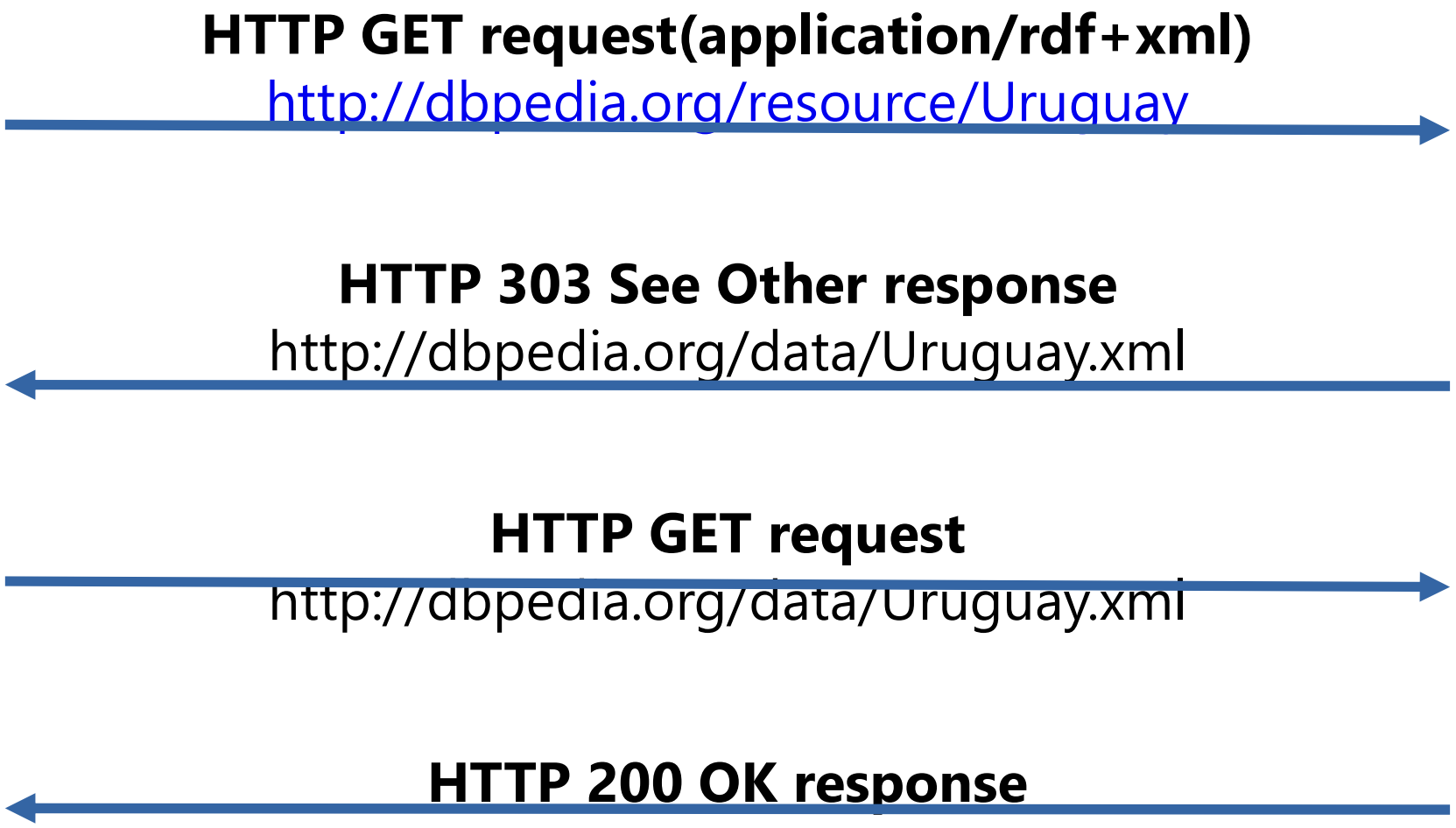

HTTP GET request(text/html)
http://dbpedia.org/resource/Uruguay
HTTP 303 See Other response
http://dbpedia.org/page/Uruguay
HTTP GET request
http://dbpedia.org/page/Uruguay
HTTP 200 OK response
Código HTML

curl -i -L -H **"Accept: text/html"** **http://dbpedia.org/resource/Uruguay**

HTTP/1.1 303 See Other
Date: Wed, 02 Oct 2013 22:21:01 GMT
Content-Type: text/html; charset=UTF-8
Server: Virtuoso/07.00.3203 (Linux) i686-generic-linux-glibc212-64 VDB
Location: http://dbpedia.org/page/Uruguay
...

HTTP/1.1 200 OK
Date: Wed, 02 Oct 2013 21:59:40 GMT
Content-Type: text/html; charset=UTF-8
Content-Length: 608380
Server: Virtuoso/07.00.3203 (Linux) i686-generic-linux-glibc212-64 VDB
...

<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" ...


HTTP GET request(application/rdf+xml)
http://dbpedia.org/resource/Uruguay
HTTP 303 See Other response
http://dbpedia.org/data/Uruguay.xml
HTTP GET request
http://dbpedia.org/data/Uruguay.xml
HTTP 200 OK response
Datos en RDF

curl -i -L -H **"Accept: application/rdf+xml"** **http://dbpedia.org/resource/Uruguay**

HTTP/1.1 303 See Other
Date: Wed, 02 Oct 2013 22:28:35 GMT
Content-Type: application/rdf+xml; qs=0.95
Server: Virtuoso/07.00.3203 (Linux) i686-generic-linux-glibc212-64 VDB
Location: http://dbpedia.org/data/Uruguay.xml
...

HTTP/1.1 200 OK
Date: Wed, 02 Oct 2013 22:28:35 GMT
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 603067
...

<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:dcterms="http://purl.org/dc/terms/"
• • •

curl -i -L -H **"Accept: text/rdf+n3"** **http://dbpedia.org/resource/Uruguay**

HTTP/1.1 303 See Other
Date: Wed, 02 Oct 2013 22:29:03 GMT
Content-Type: text/rdf+n3; qs=0.8
Server: Virtuoso/07.00.3203 (Linux) i686-generic-linux-glibc212-64 VDB
Location: http://dbpedia.org/data/Uruguay.nt
...

HTTP/1.1 200 OK
Date: Wed, 02 Oct 2013 22:29:04 GMT
Content-Type: text/rdf+n3; charset=UTF-8
Content-Length: 296903
...

@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix ns1: <http://cs.dbpedia.org/resource/> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
ns1:Uruguay owl:sameAs dbpedia:Uruguay .
@prefix ns3: <http://pl.dbpedia.org/resource/> .
ns3:Urugwaj owl:sameAs dbpedia:Uruguay .

Utilización de Hash URIs

```
curl -i  http://www.w3.org/2006/07/SWD/recipes/examples-20080421/
example1#ClassA

HTTP/1.1 200 OK
Date: Thu, 03 Oct 2013 13:08:38 GMT
Server: Apache/2
Content-Type: application/rdf+xml;
...

<?xml version="1.0"?>
<rdf:RDF
xml:base="http://www.w3.org/2006/07/SWD/recipes/examples20080421/example1">
<owl:Ontology
rdf:about="http://www.w3.org/2006/07/SWD/recipes/examples20080421/example1">
  <rdfs:comment>An example vocab for minimal hash setup.</rdfs:comment>
  <rdfs:label>Example 1 Ontology</rdfs:label>
</owl:Ontology>
<owl:Class rdf:about="http://www.w3.org/2006/07/SWD/recipes/examples-
20080421/example1#ClassA">
</owl:Class> ...
</rdf:RDF>
```

3) Proveer información **útil** acerca de cada URI en RDF

```
dbpedia:Eladio_Dieste       dbpedia-owl:birthPlace dbpedia:Uruguay ;
        dbpprop:birthPlace          dbpedia:Uruguay ;
        dbpprop:deathPlace          dbpedia:Uruguay ;
        dbpedia-owl:deathPlace      dbpedia:Uruguay .
```

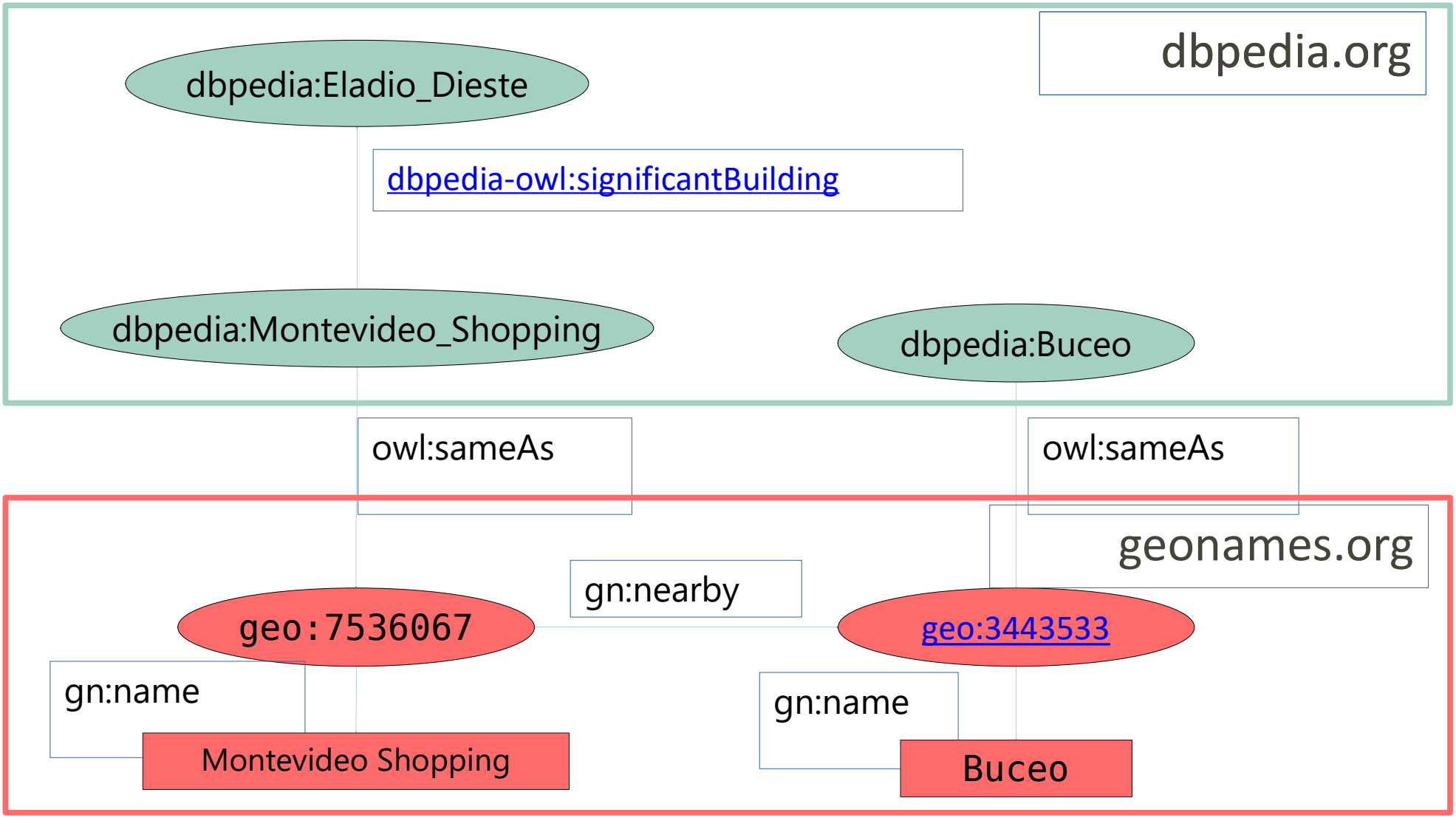
4) Crear links entre URIs

http://dbpedia.org/resource/Uruguay

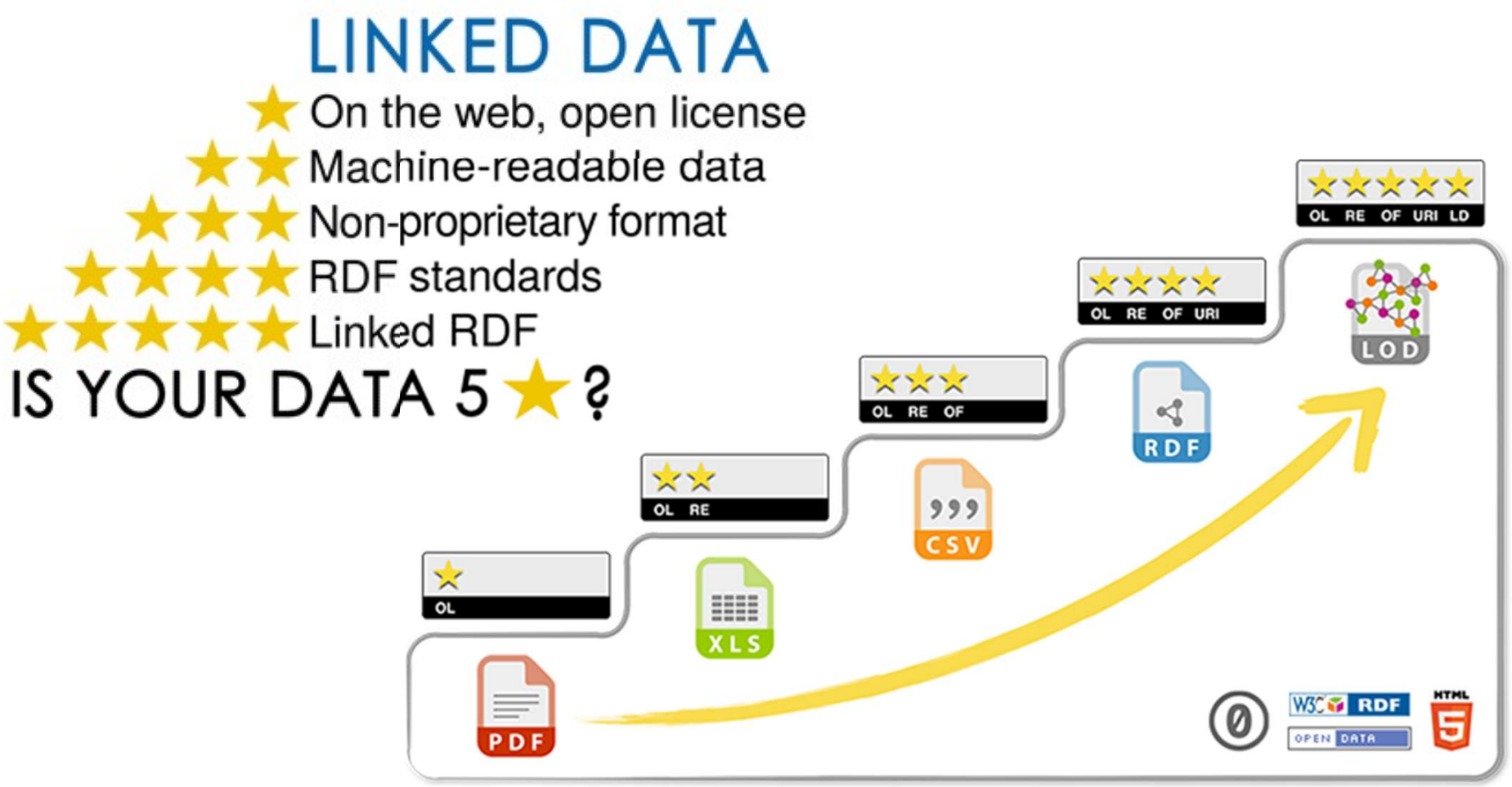
```
@prefix ns192:   <http://www.wikidata.org/entity/> .
@prefix ns17:    <http://nl.dbpedia.org/resource/> .
@prefix ns18:    <http://eu.dbpedia.org/resource/> .

dbpedia:Uruguay          owl:sameAs       ns192:Q77 ,
                         <http://gadm.geovocab.org/id/0_233> ,
                         <http://rdf.freebase.com/ns/m.07twz> ,
                         ns18:Uruguai ,
                         ns17:Uruguay .
```

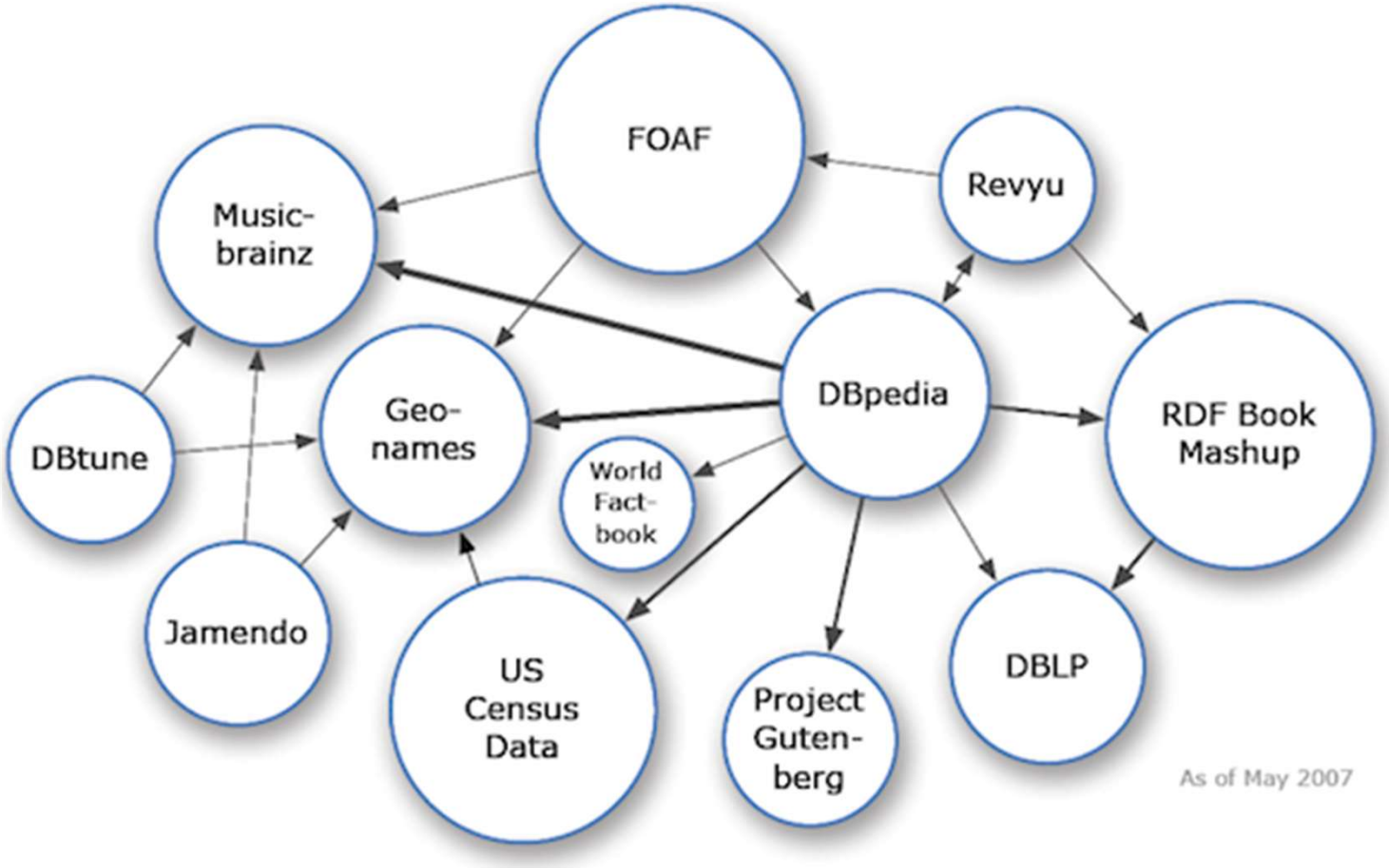
```
@prefix gn: http://www.geonames.org/ontology
@prefix dbpedia: http://dbpedia.org/resource/
@prefix dbpedia-owl: http://dbpedia.org/ontology/
@prefix geo:http://sws.geonames.org/
@prefix owl:http://www.w3.org/2002/07/owl#
```



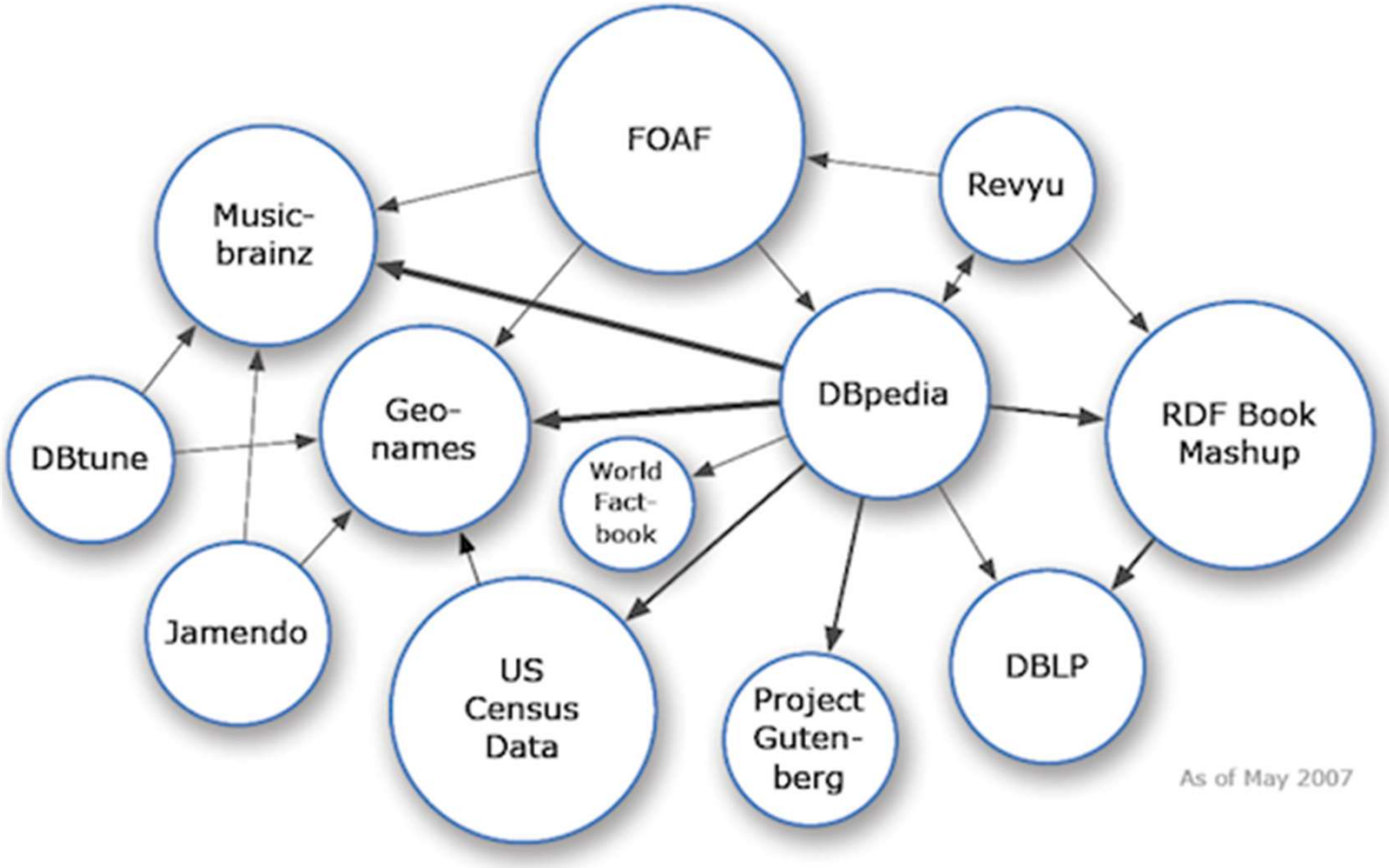
Linked Data + Open Data
= LOD



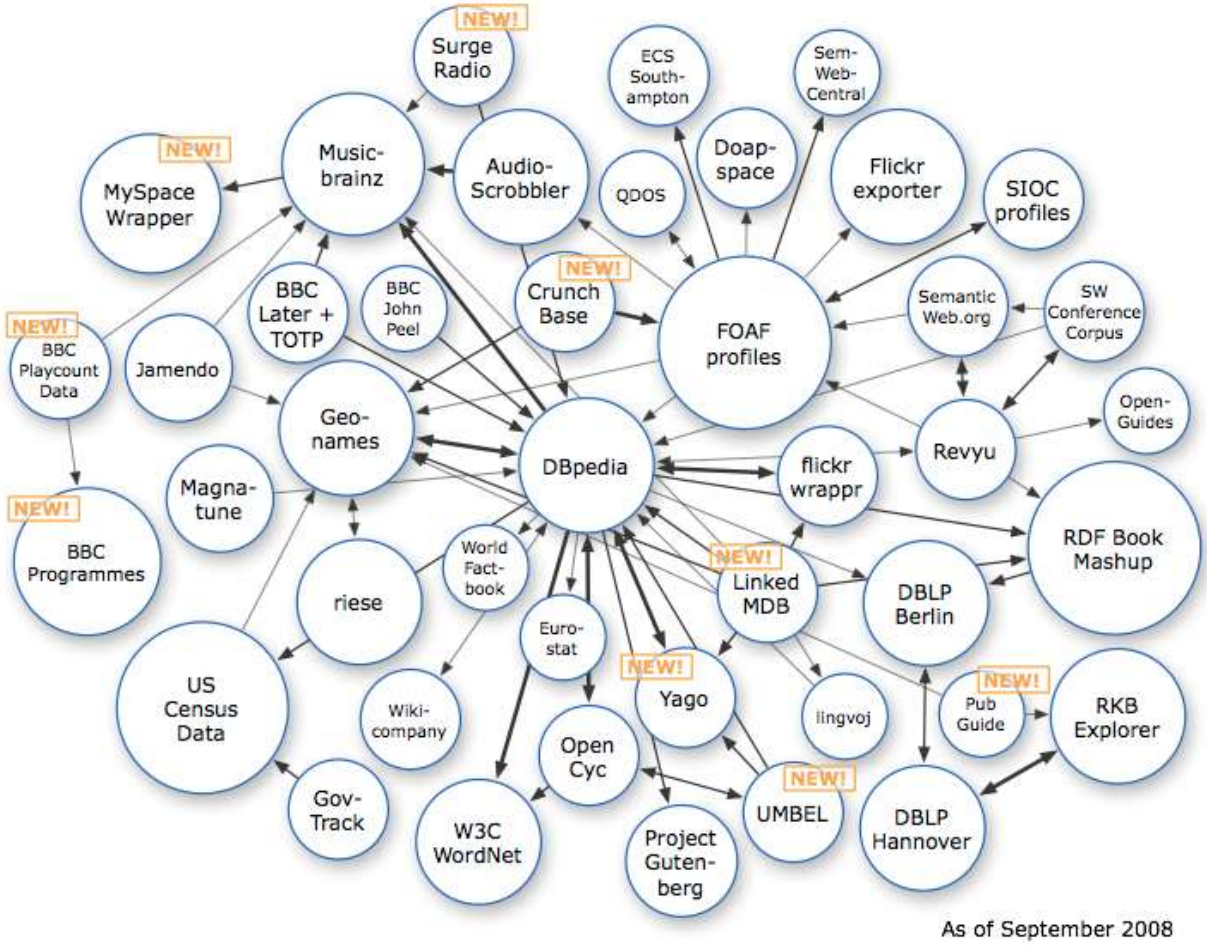
http://5stardata.info/en/



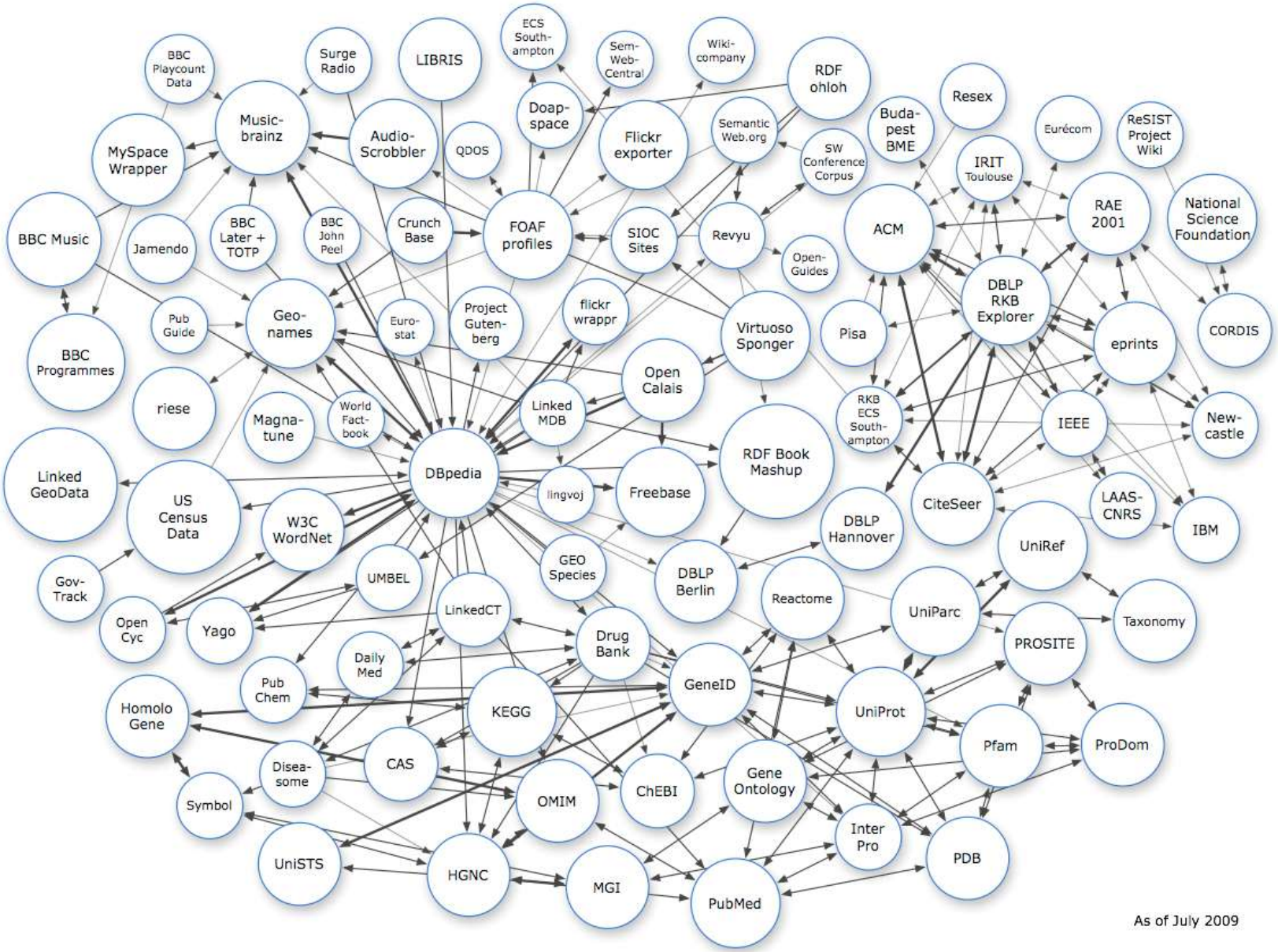
http://lod-cloud.net/ 2007



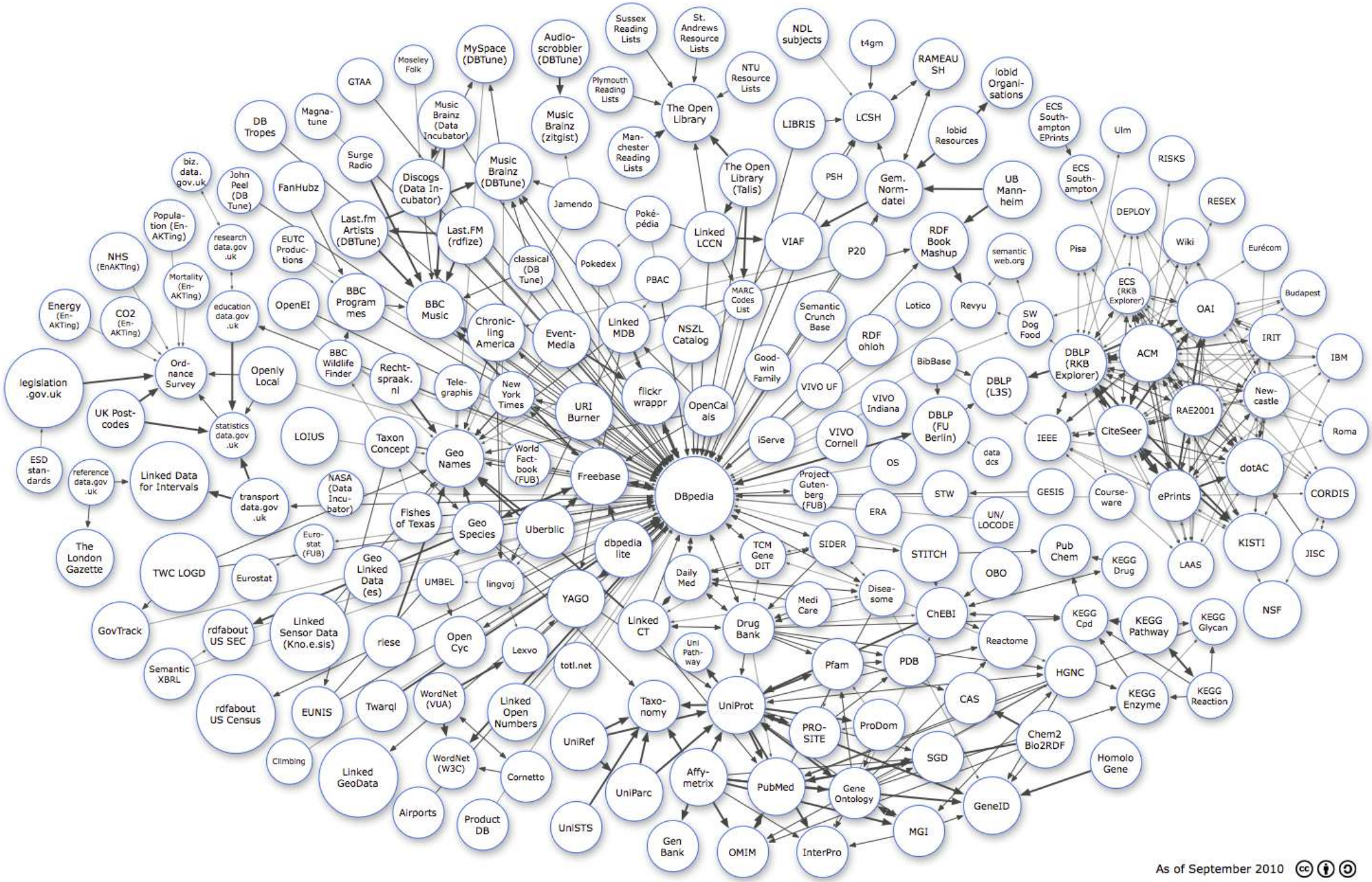
http://lod-cloud.net/ 2007



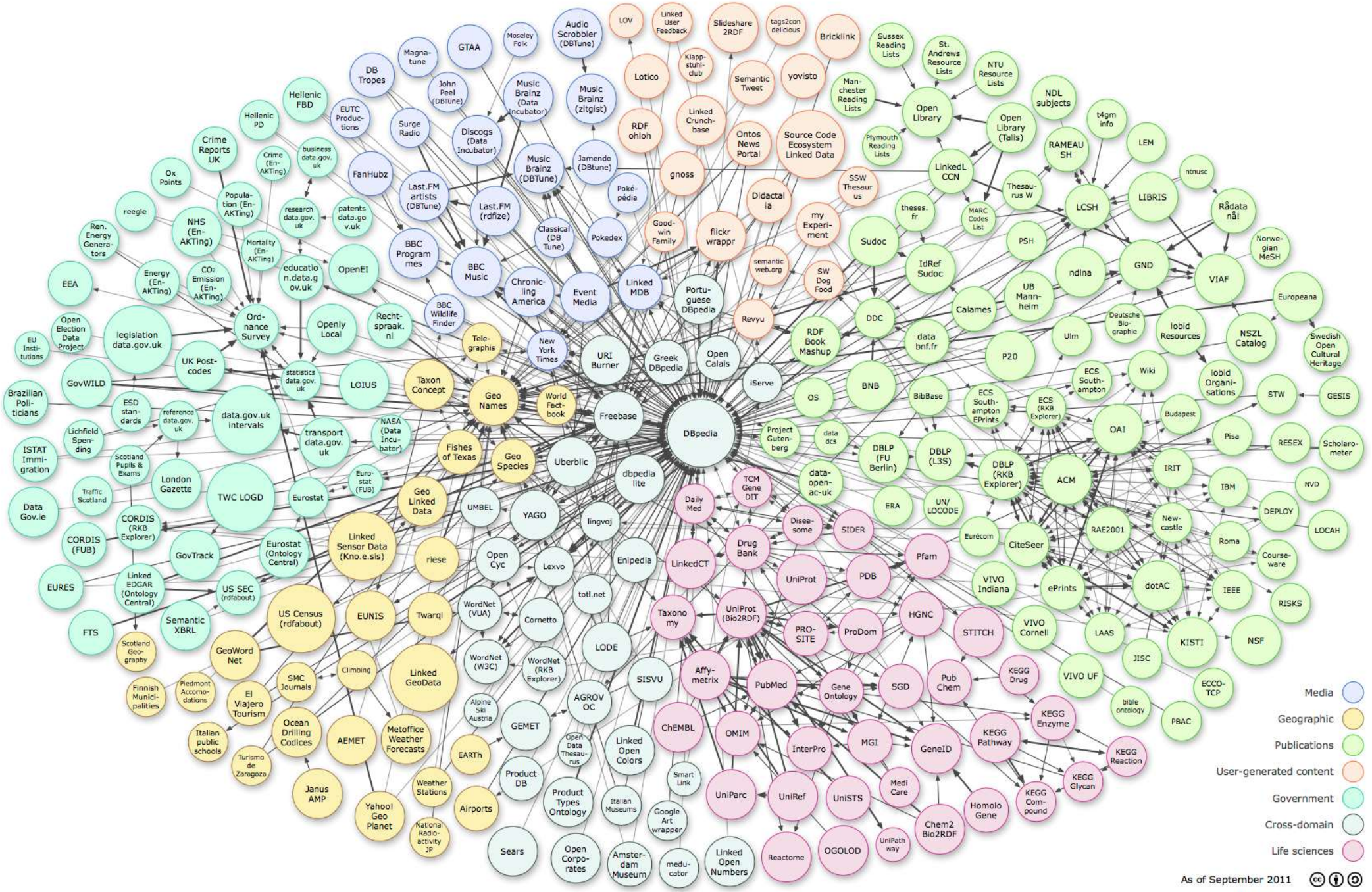
http://lod-cloud.net/ 2008



http://lod-cloud.net/ *2009*



http://lod-cloud.net/ 2010



http://lod-cloud.net/ *2011*

Algunos casos interesantes usando PGM

Analyzing the Panama Papers with Neo4j: Data Models, Queries & More
https://neo4j.com/blog/analyzing-panama-papers-neo4j/

Analyzing the Graph of Thrones
http://www.lyonwj.com/2016/06/26/graph-of-thrones-neo4j-social-network-analysis/