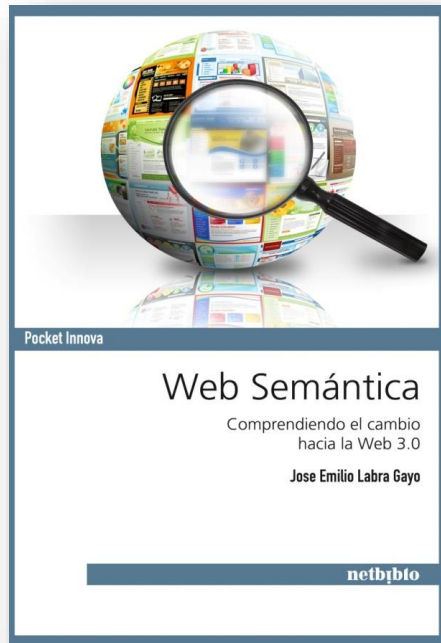


Introducción a la Web Semántica

Jose Emilio Labra Gayo

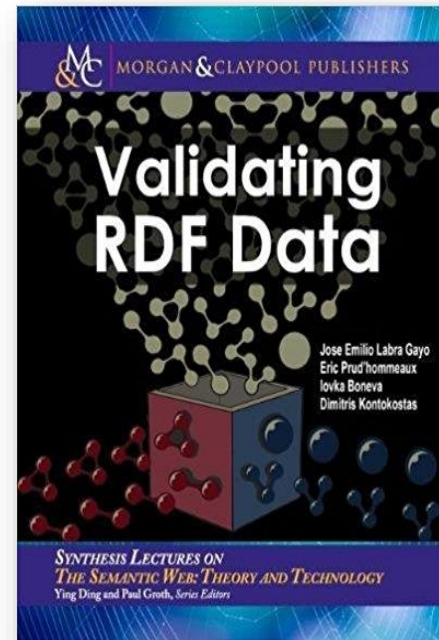
Departamento de Informática
Universidad de Oviedo

1' de publicidad



Web Semántica

Editorial NetBiblo, Colección Pcket Innova, 2012
<http://www.netbiblo.com>



Validating RDF Data

Ed. Morgan & Claypool, 2018
<http://book.validatingrdf.com>
Online HTML version

Web semántica

Visión de la Web como una **web de datos**

No solo páginas web, sino datos

Datos enlazados

Campo relacionado con:

Big Data

Enormes cantidades de datos de la Web
...¡y más datos que se van a generar!

Inteligencia Artificial

Representación del conocimiento
Inferencia de nuevo conocimiento

Ejemplo: Wikipedia/Wikidata



Tim Berners-Lee
Fuente: Wikipedia

1,677,782,739

Websites online right now

Fuente: <http://www.internetlivestats.com/total-number-of-websites/>

Consultado: 05/04/2019 (19:05h)

¿Quién consume información de la Web?

Personas

Accedemos a través de un dispositivo

...y **Máquinas** (programas informáticos)

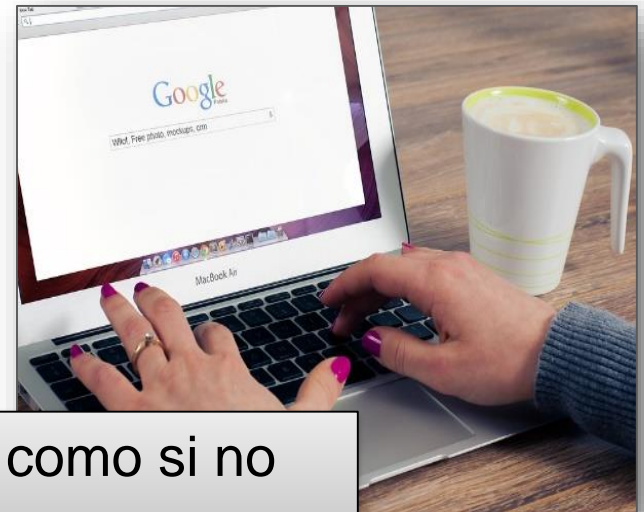
Nos muestran las páginas Web (navegadores)

...pero también analizan la información (bots)

Nos filtran contenido

Nos hacen sugerencias

...



"Si Google no *entiende* tu página Web es como si no existiese"

¿Personas vs Máquinas?



Creatividad, imaginación
Imprevisibles (cometemos errores)
Nos cansamos ante tareas repetitivas
Comprensión basada en contexto



Programadas para ciertas tareas
Previsibles (sin errores*)
Tareas repetitivas sin problema
Dificultad para entender el contexto

*cuando están bien programadas

¿Información *entendible* por las máquinas?

Problema: Ambigüedad e identificación del contexto

Ejemplo: "Oviedo tiene una temperatura de 36 grados"

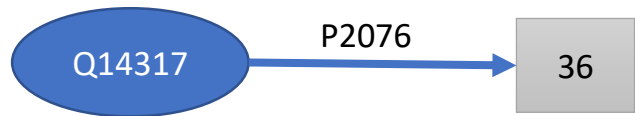
¿ **Oviedo** ? Puede ser: Una ciudad en España
...o una ciudad en Florida
...o un jugador de fútbol

...tiene una temperatura de...

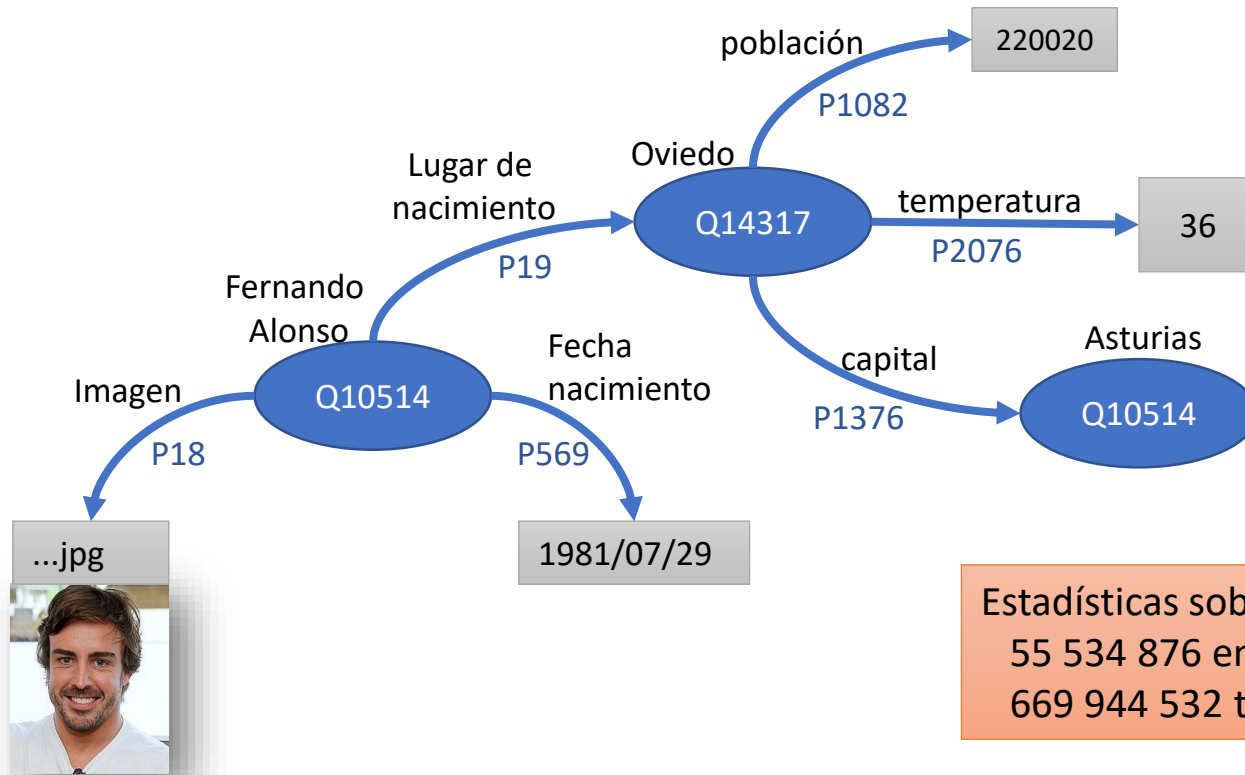
Identificadores (URIs) en Wikidata
<http://wikidata.org/entity/Q14317>
<http://www.wikidata.org/entity/Q1813449>
<http://www.wikidata.org/entity/Q325997>

<https://www.wikidata.org/wiki/Property:P2076>

Representación para máquinas



Grafos de conocimiento



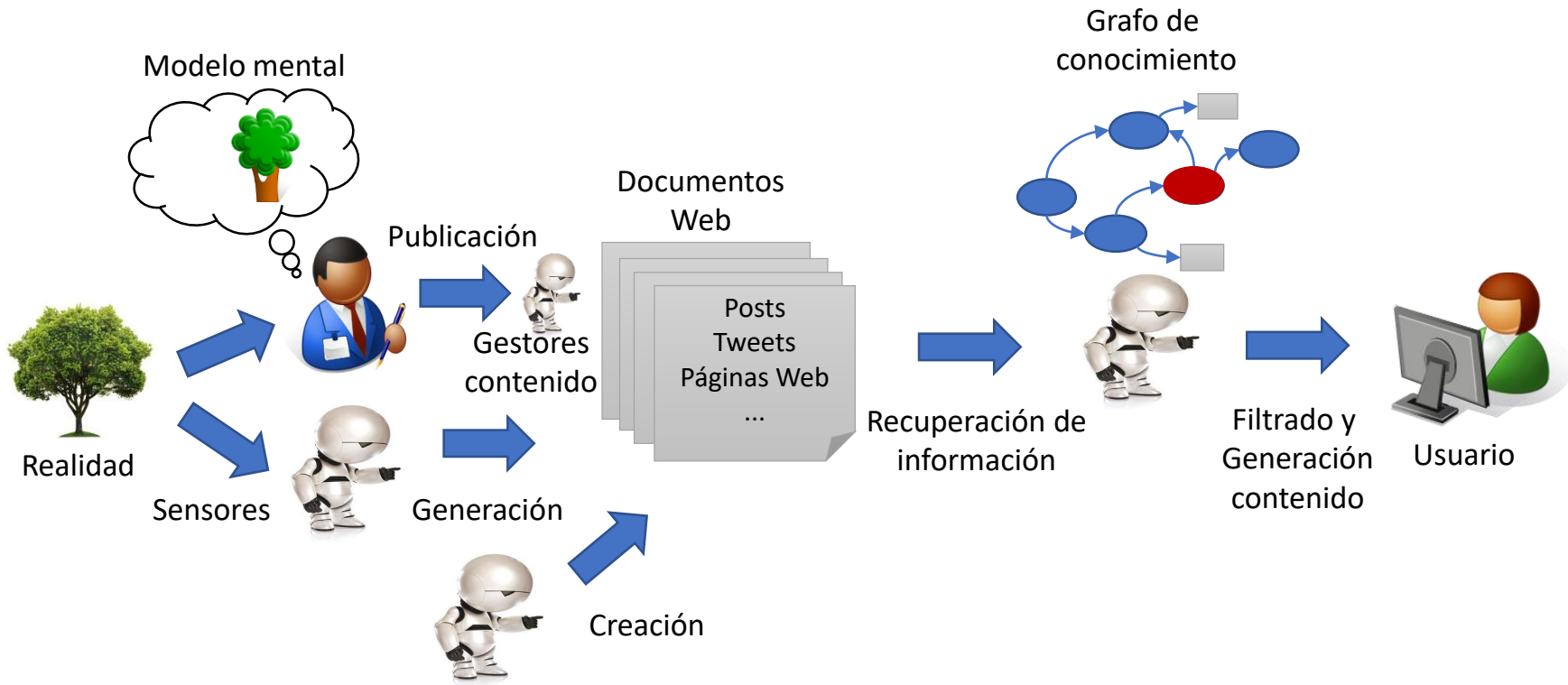
Estadísticas sobre Wikidata (25-03-2019)
55 534 876 entidades
669 944 532 tripletas

Wikidata es un ejemplo de grafo de conocimiento
...pero hay muchos más...

Abiertos: DBpedia, BabelNet, ...

Propietarios: Google, facebook, Microsoft, etc. tienen grafos de conocimiento

Máquinas en la Web y grafos de conocimiento



Información en la Web manipulada constantemente por máquinas
Web Semántica \Rightarrow facilitar esa manipulación

Características de la web

No centralizada

Difícil garantizar integridad de la información

Información Dinámica

La información existente cambia

Mucha información

Un sistema no puede pretender acaparar toda la información

Es abierta

Open World Assumption

Principio **CCC**

Cualquiera puede decir **Cualquier** cosa sobre **Cualquier** tema

En inglés: Principio **AAA**: Anyone can say Anything about Any topic
Fuente: Semantic Web for the Working Ontologist, D. Allemang, J. Hendler

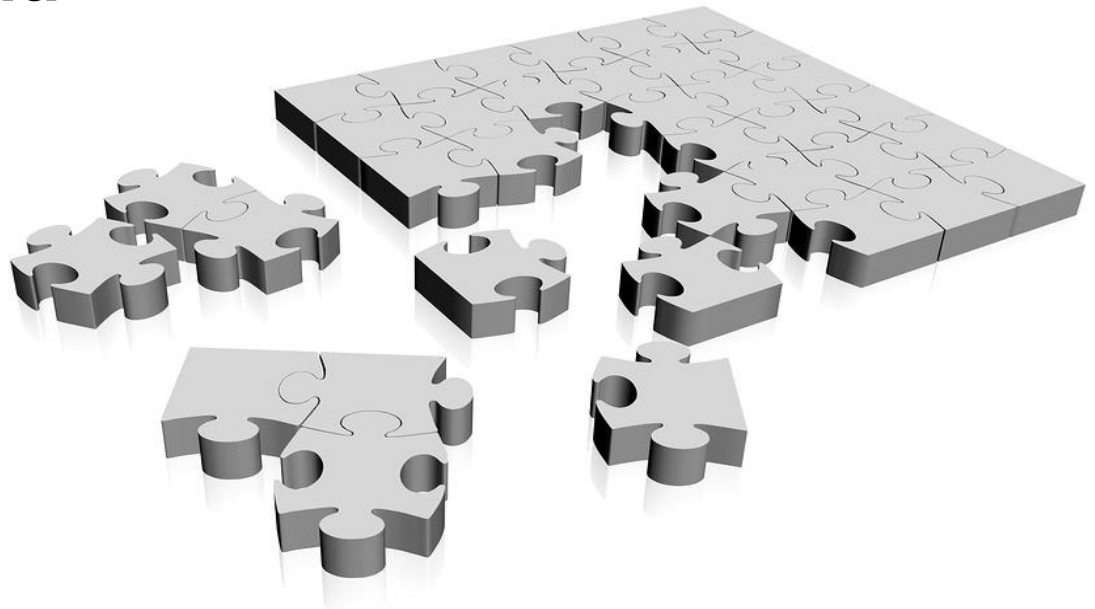
No basta con publicar datos...

El mayor reto = Integración

En general, el problema *no es informatizar* algo

El problema es **integrar** los sistemas

Interoperabilidad



Modelo de Estrellas*

- ★ **Publicar** los datos
(en cualquier formato)
- ★★ Utilizar **formato estructurado**
(Excel en lugar de imágenes escaneadas)
- ★★★ Usar formatos **no propietarios**
(CSV en lugar de Excel)
- ★★★★ Usar **URIs para identificar** datos
(otros sistemas puedan enlazar nuestros datos)
- ★★★★★ **Enlazar con otros** datos externos
(proporcionar contexto)

<http://5stardata.info/>

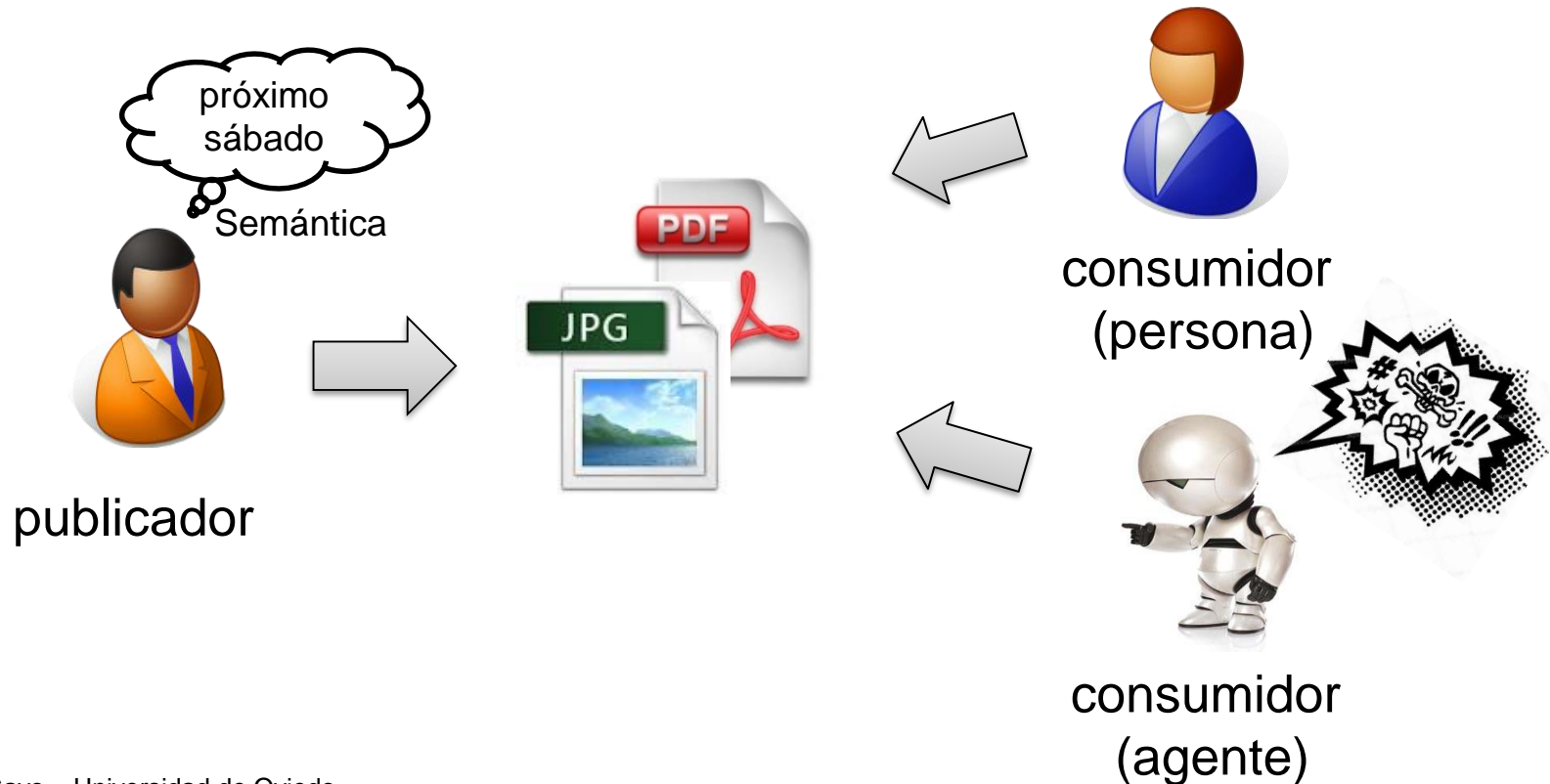
* Enunciado por Tim Berners-Lee en Gov 2.0 Expo 2010

Problema de la pérdida semántica

Pérdida de semántica en el proceso de publicación

La persona que va a publicar tiene más información

Esa información se pierde en el proceso





Formatos estructurados

Los datos tienen una estructura

Ejemplo: Hojas de cálculo

Problema con formatos propietarios

Requieren herramientas que no son públicas

No resuelven el problema de la pérdida semántica

¿Qué significa cada celda?





Formatos no propietarios

Utilizar formatos abiertos estructurados

Ejemplos: CSV, HTML, JSON, XML...

Permiten procesamiento automático

Pero no resuelven el problema de la pérdida semántica

El problema de HTML



HTML tiene como objetivo publicar hipertexto

Etiquetas HTML legibles por los navegadores

Información dentro de marcas = lenguaje natural

Las máquinas no entienden el lenguaje natural

```
<p>Evento:  
<ul>  
<li>Nombre: Concierto</li>  
<li>Fecha: Próximo sábado</li>  
</ul>  
</p>
```

```
<p>իրադարձություն:  
<ul>  
<li>տիպ: համերգ</li>  
<li>ամսաթիվ: հաջորդ շաբաթ</li>  
</ul>  
</p>
```


El problema de XML



XML da un paso más hacia la solución

Se pueden definir vocabularios específicos

Pueden crearse aplicaciones que los procesan

Sin embargo, los documentos XML no se integran fácilmente si son de otros vocabularios

```
<event>  
  <name>Concierto</name>  
  <date>Próximo sábado</date>  
</event>
```

```
<event>  
  <name>համերգ</name>  
  <date>հաջորդ շաբաթ</date>  
</event>
```

```
<իրադարձություն>  
  <տիպ>համերգ </տիպ>  
  <ամսաթիվ>հաջորդ շաբաթ</ամսաթիվ >  
</իրադարձություն>
```

¿Y JSON?



Más o menos...lo mismo que XML

JSON tiene un modelo jerárquico similar a XML

Aunque existe JSON Schema, la validación es menos habitual

Los nombres de los campos son cadenas de texto

```
{ "event":  
  { "name": "Concierto" ,  
    "date": "Próximo sábado"  
  }  
}
```

```
{ "event":  
  { "name": "համերգ" ,  
    "date": "հաջորդ շաբաթ"  
  }  
}
```

```
{ "իրադարձություն":  
  { "տիպ": "համերգ" ,  
    "ամսաթիվ": "հաջորդ շաբաթ"  
  }  
}
```



URIs para identificar datos

Utilizar URIs para identificar datos

Negociación de contenido

Devolver diferentes representaciones

Ejemplo:

HTML para personas con navegadores

RDF para sistemas automáticos



Ejemplo: RDF



<<http://www.sepe.es/datos/desempleo/Asturias/Allande/2013/10>>



¿Varias representaciones de lo mismo?

La arquitectura de la web separa recurso de representación

Ejemplo: Bolsa de patatas fritas



Enlazar con otros datos

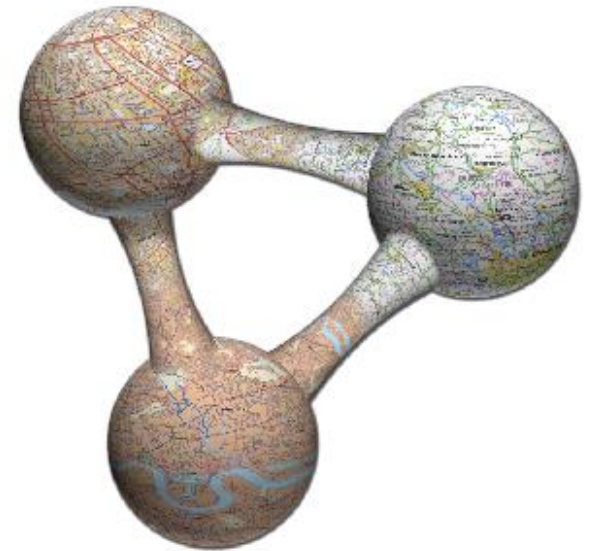


Las representaciones devueltas incluyen enlaces con otros datos

Permite:

Reutilizar y descubrir datos

Aplicaciones "*no previstas*"



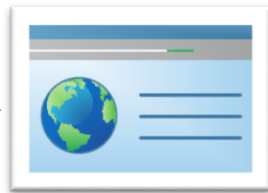
Ejemplo: RDF bien enlazado



<<http://www.sepe.es/datos/desempleo/Asturias/Allende/2013/10>>

RDF?

HTML?



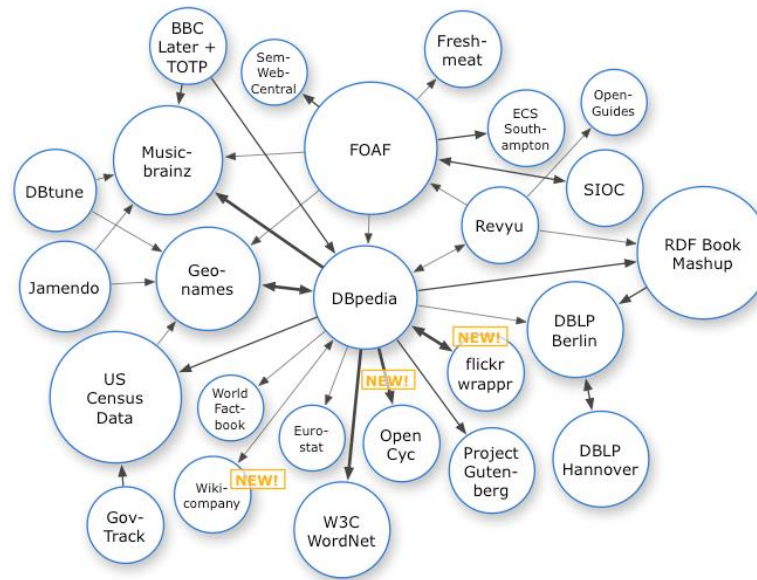
```
@prefix sepe: <http://www.sepe.es/datos/>
sepe:obs1 sepe:municipio <http://dbpedia.org/resource/Allende>;
          sepe:desempleados 23 .
```

```
dbo:allende dbo:areaTotal 342.24 ;
            rdf:type <http://.../municipalitiesInAsturias> ;
            dbo:country <http://.../Spain> ;
            dbo:populationTotal 2106 ;
            . . .
```

Principios Linked Open Data

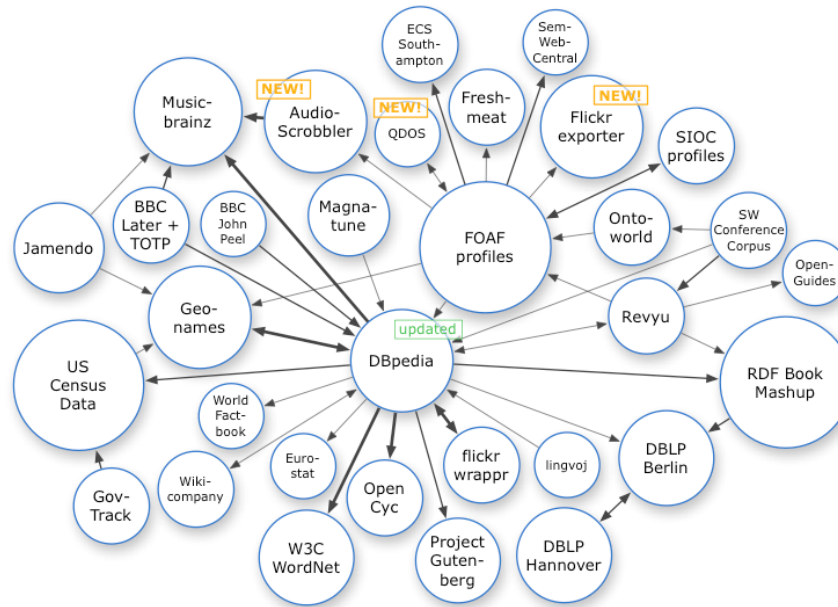
1. Utilizar URIs para denotar cosas
2. Permitir que las URIs sean dereferenciables
3. Proporcionar información útil
Para personas y máquinas (HTML, RDF)
4. Incluir enlaces a otras cosas relacionadas

LOD (2007)

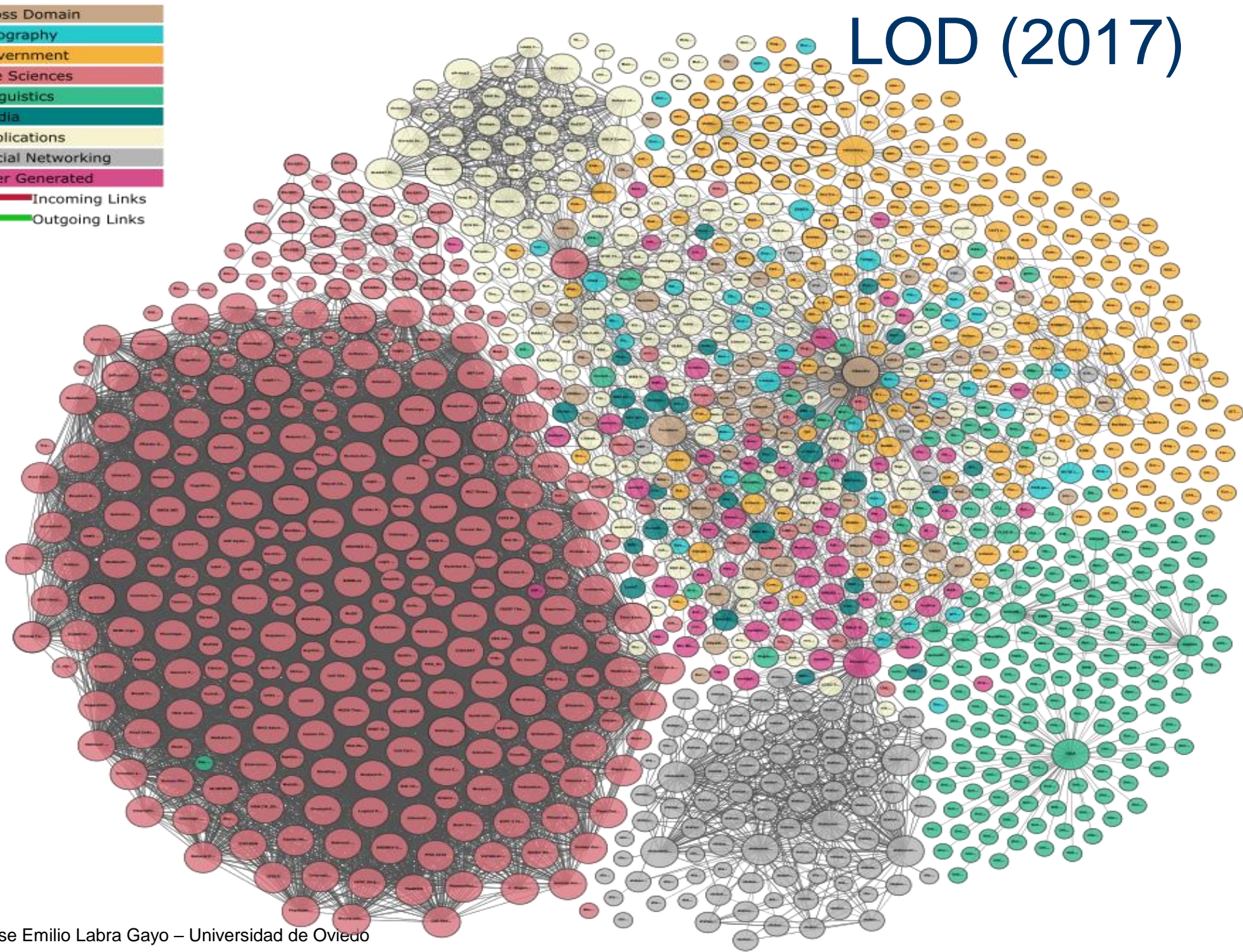


Evolución: <https://lod-cloud.net/>

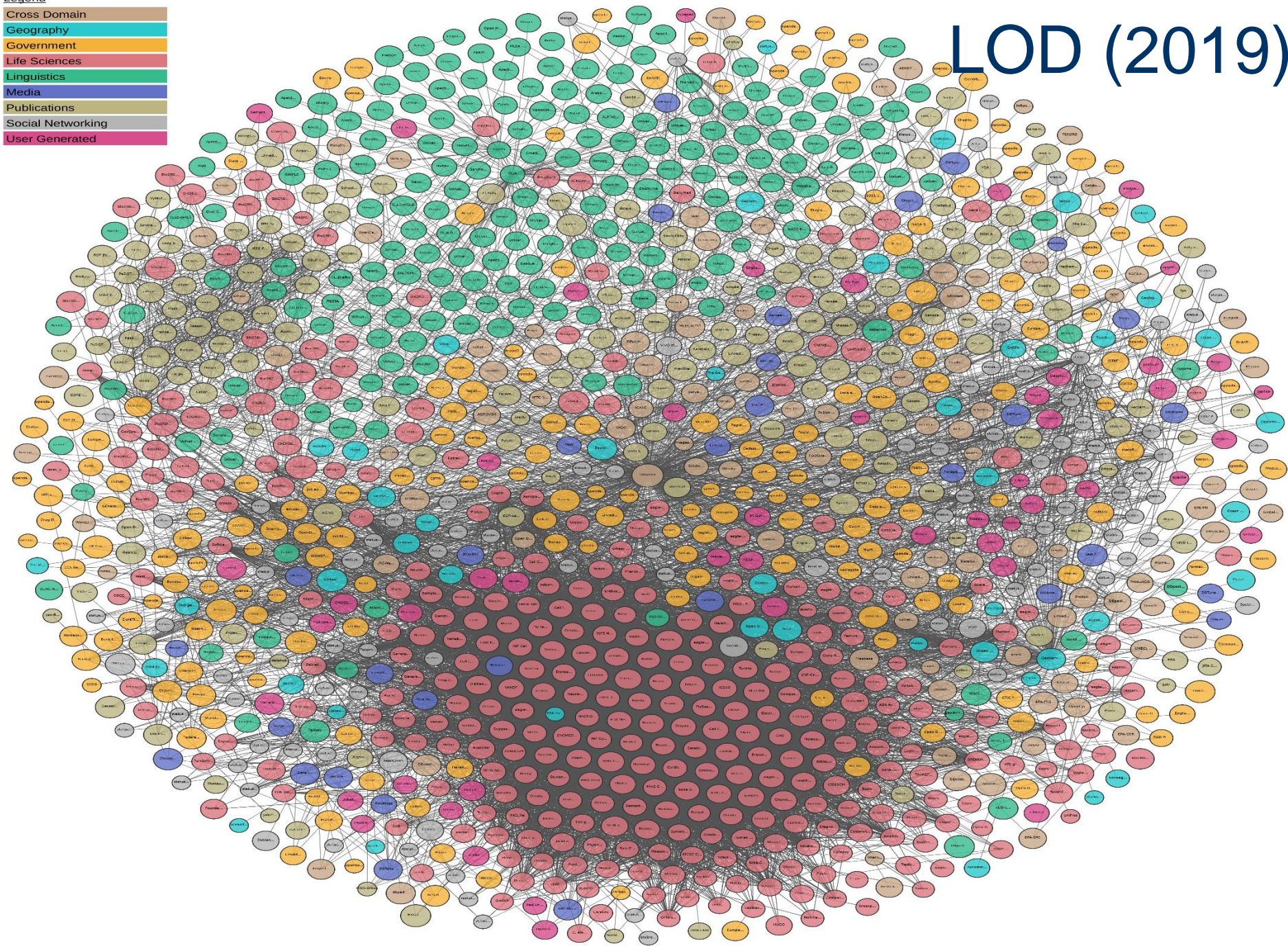
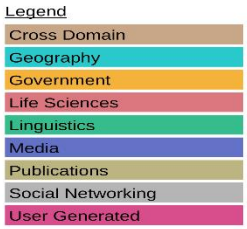
LOD (2008)



LOD (2017)



LOD (2019)



Datos abiertos enlazados

Ejemplos de iniciativas

data.gov.uk

data.worldbank.org

data.gov

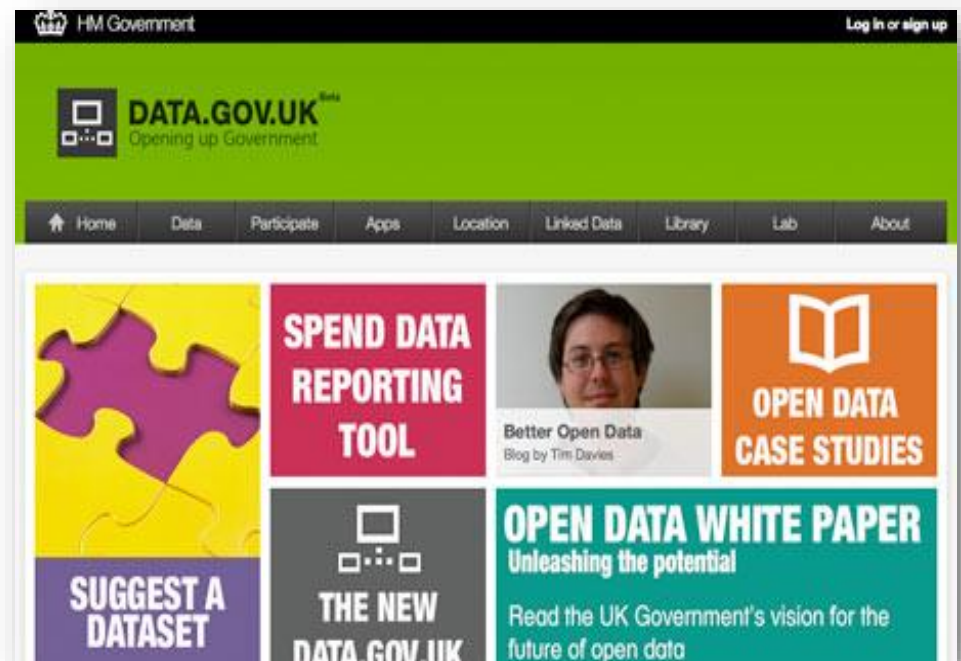
datos.gob.es

datos.gijon.es

...

datos.bcn.cl

data.webfoundation.org



Beneficios de Linked Open Data

Datos accesibles

Evitar pérdidas semánticas al publicar

Facilitar automatización de tareas

Datos enlazados

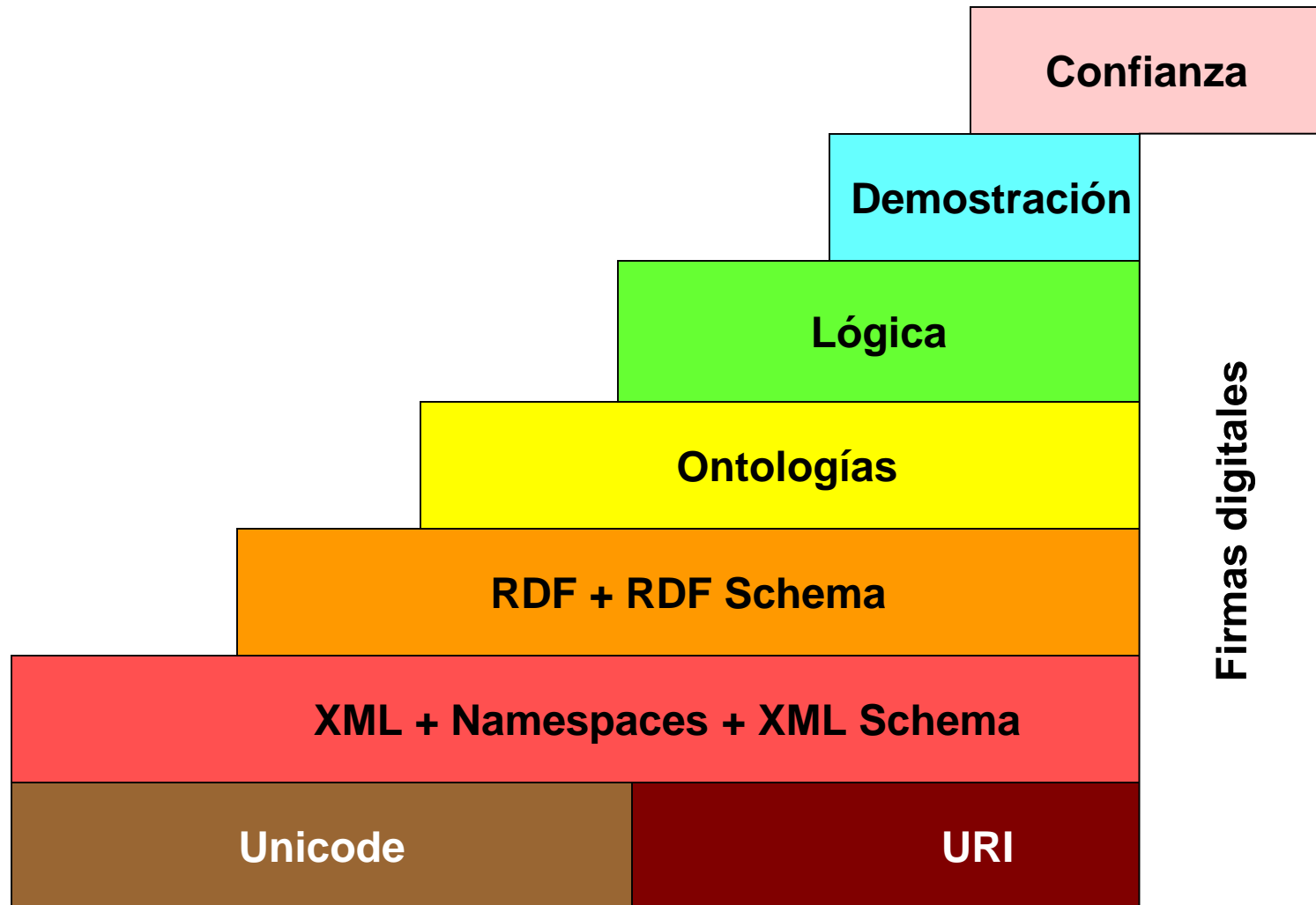
Reutilización de datos

Integración de aplicaciones

La mejor manera de explotar tus
datos se le ocurrirá a otro

Tecnologías Web Semántica

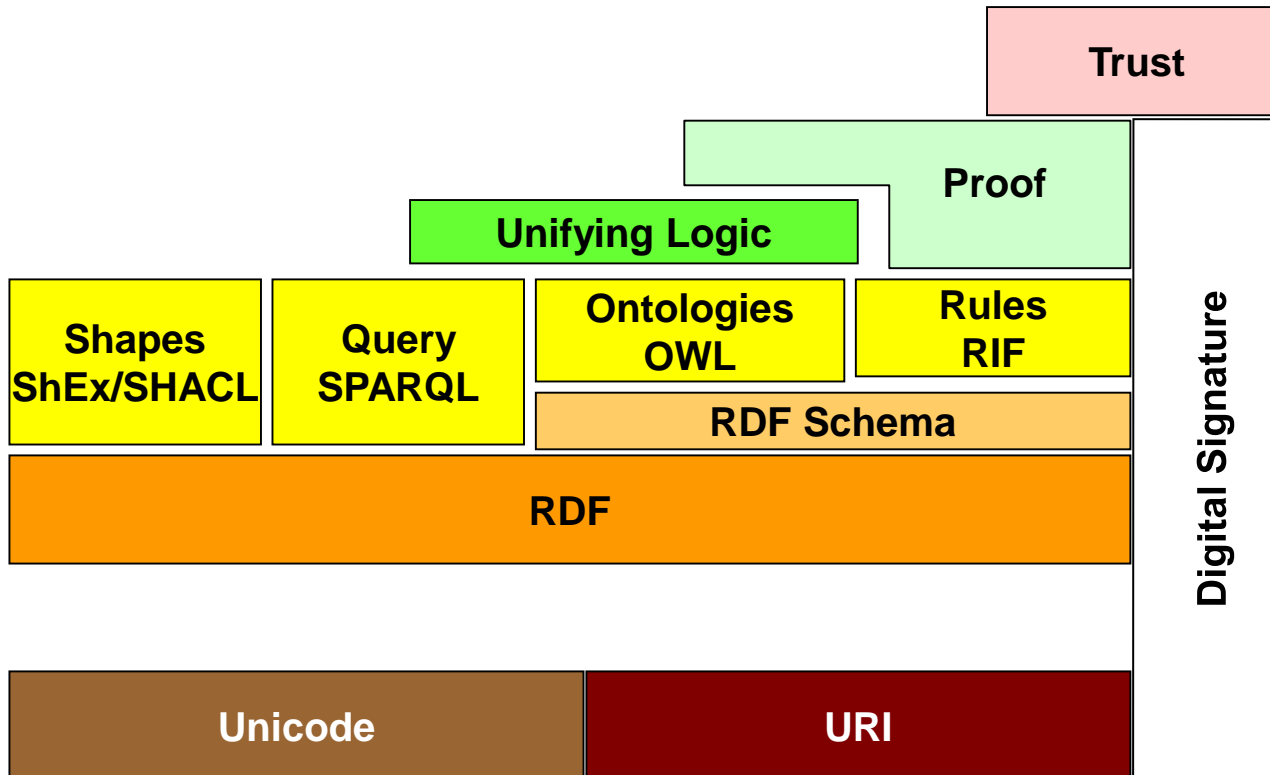
Tarta de la Web (versión inicial)



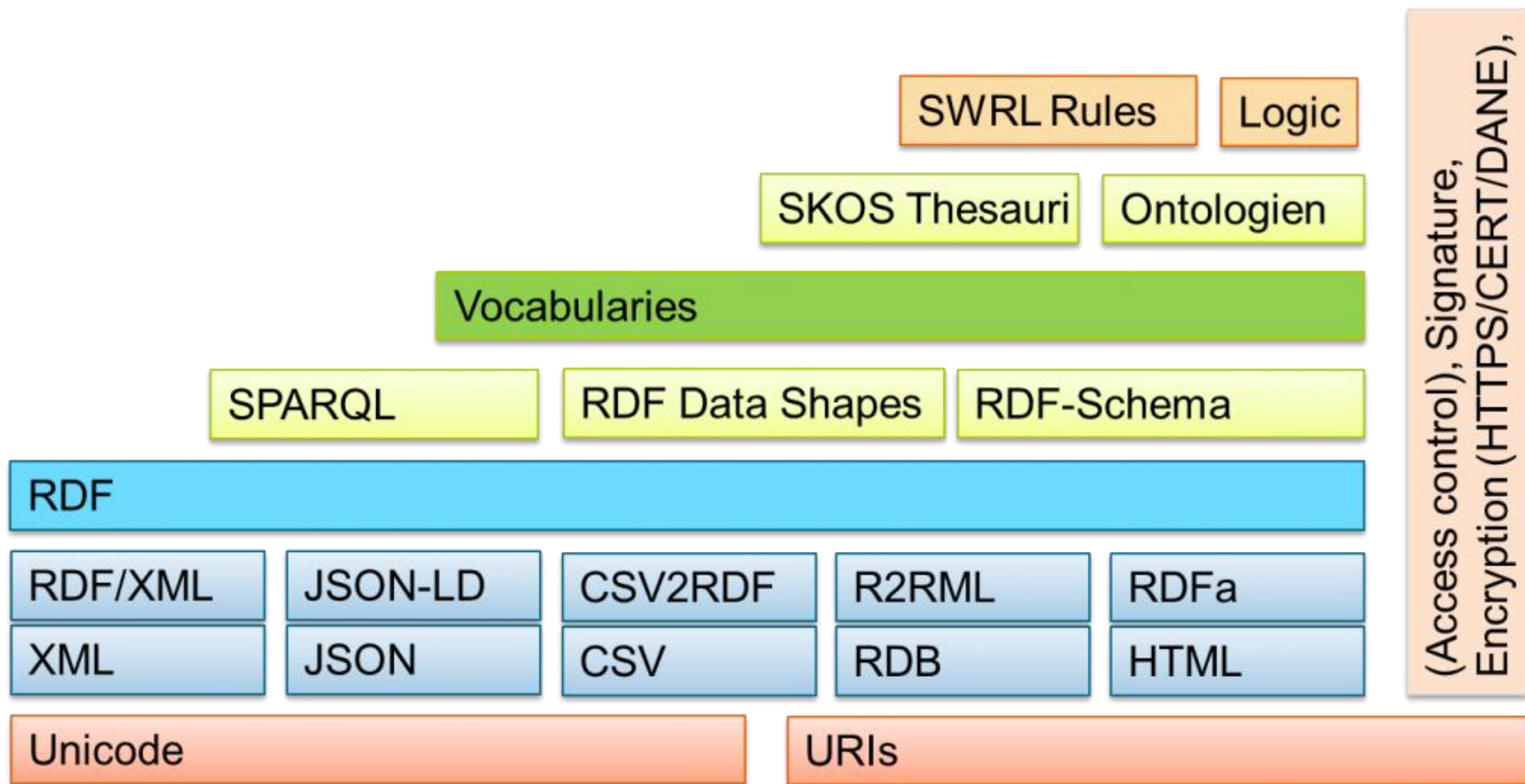
Versión propuesta por Tim Berners Lee, año 2000
<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>

Tecnologías Web Semántica

Cambios en la tarta...



Nuevas versiones de la tarta



Algunas tecnologías

RDF
Descripción datos

SPARQL
Consultas

SHEX - SHACL
Validación

OWL - RDFS
Inferencias

RDF



Resource Description Framework (1998)

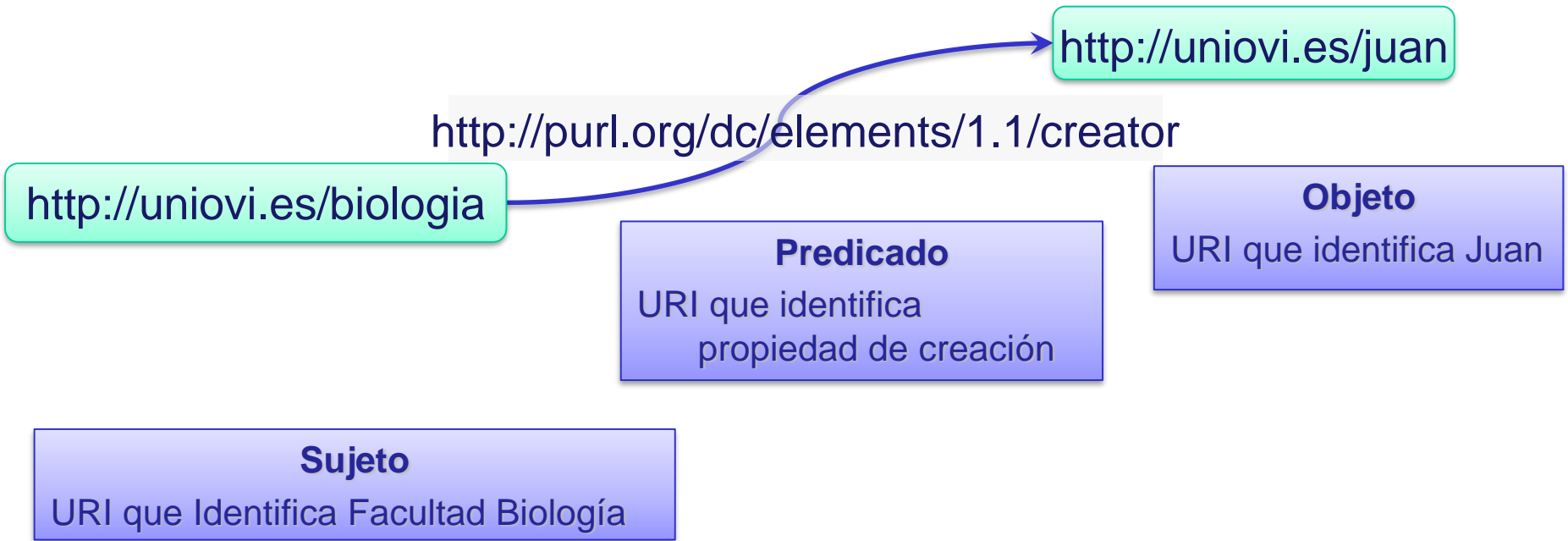
Descripción de recursos

Recurso = se identifica con URI

Se basa en tripletas

Sujeto → Predicado → Objeto

Tripletas RDF

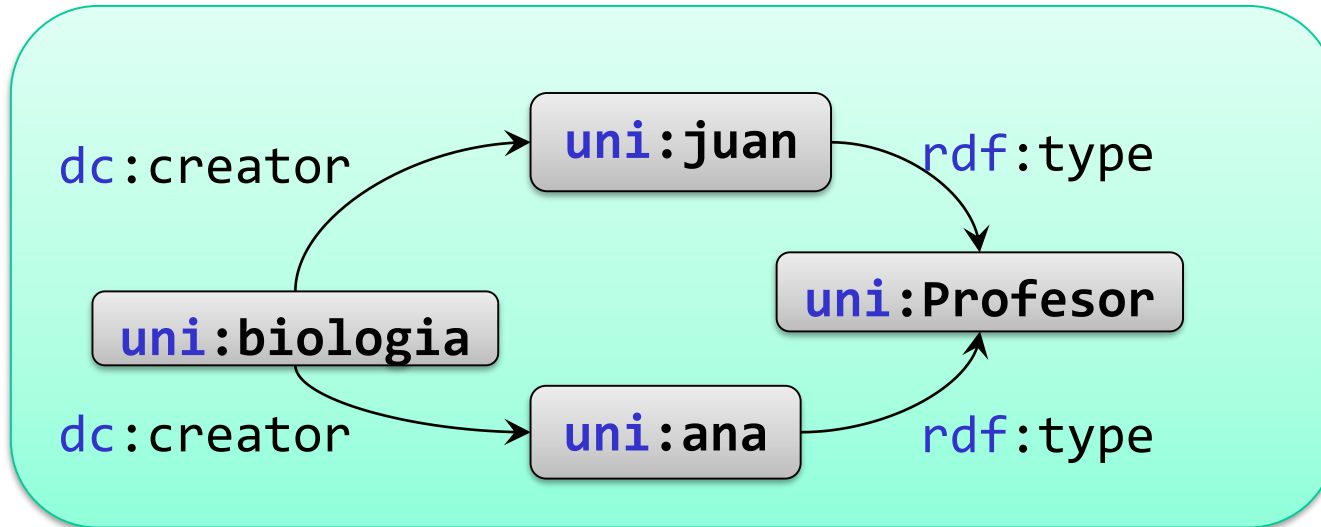


RDF en notación Turtle

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix uni: <http://uniovi.es/> .

uni:biologia dc:creator uni:juan .
```

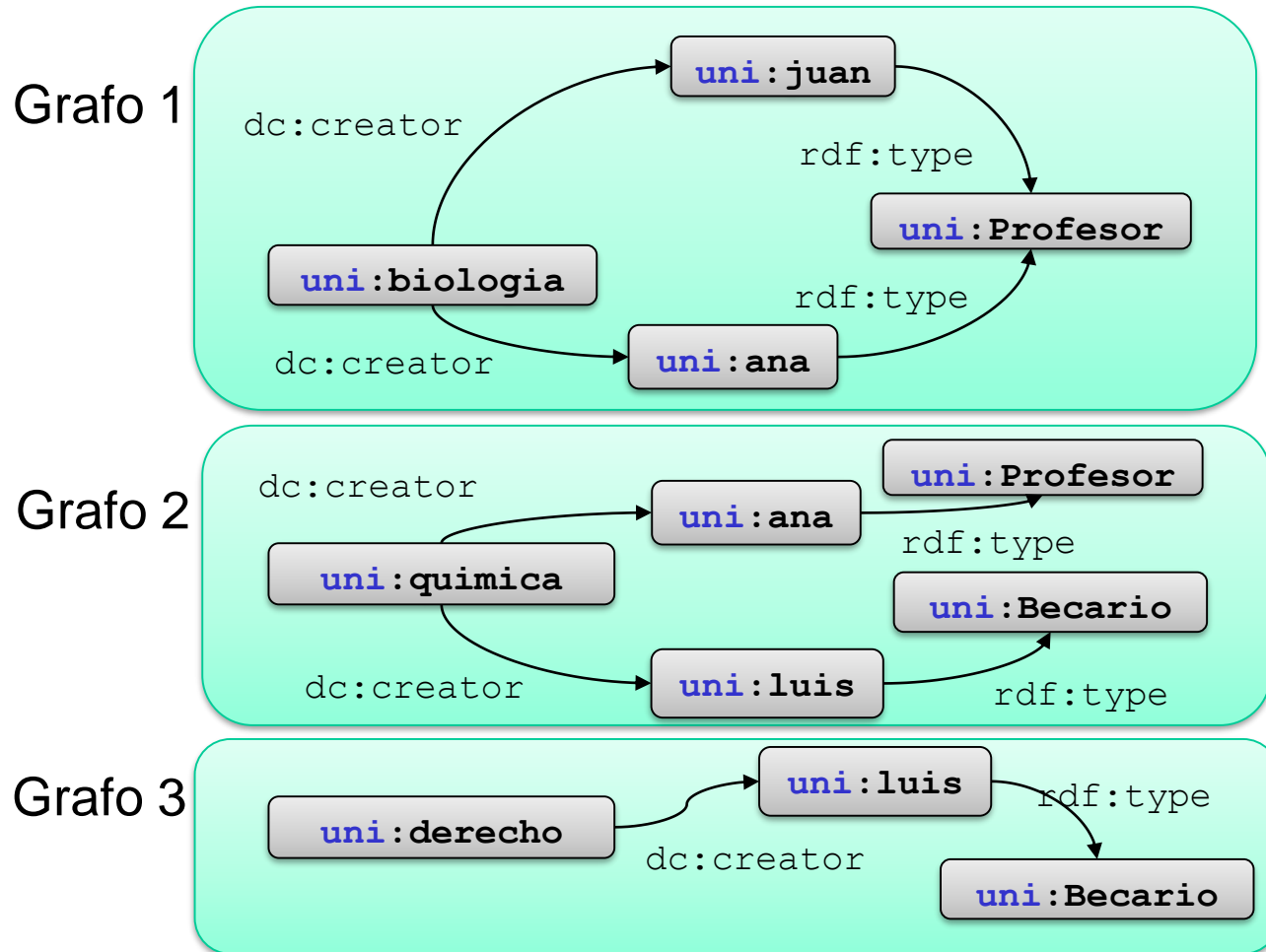

Grafo RDF



```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .  
@prefix uni: <http://uniovi.es/> .  
@prefix dc: <http://purl.org/dc/elements/1.1/> .
```

```
uni:biologia    dc:creator    uni:juan .  
uni:biologia    dc:creator    uni:ana .  
uni:juan        rdf:type      uni:Profesor .  
uni:ana         rdf:type      uni:Profesor .
```

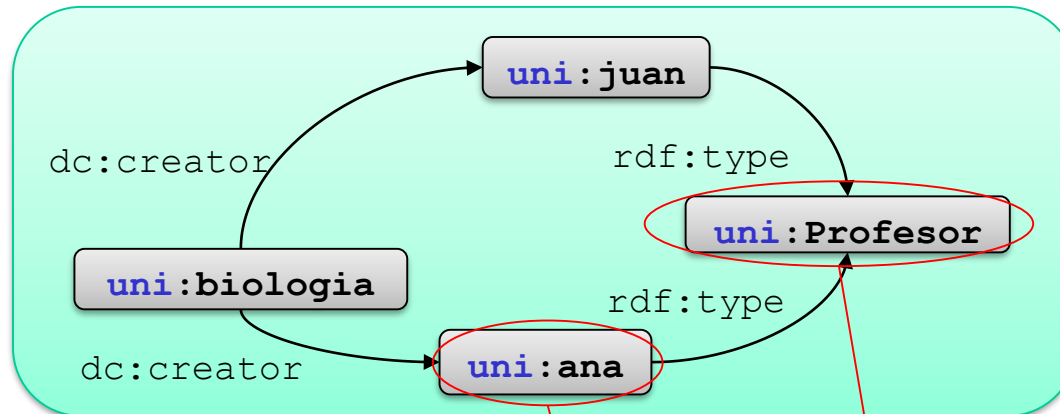
RDF es composicional



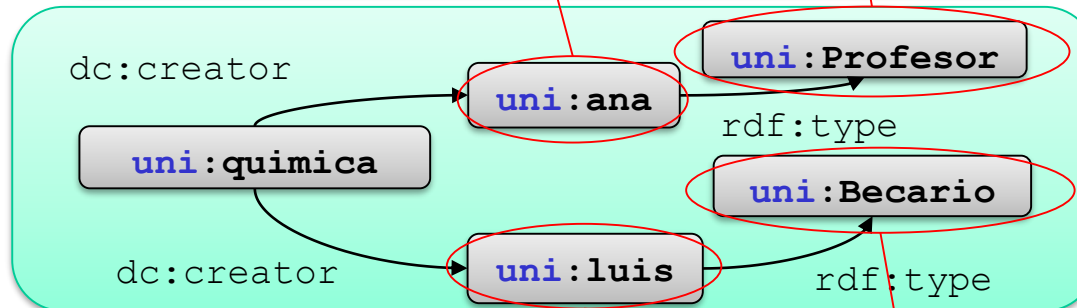
RDF es composicional



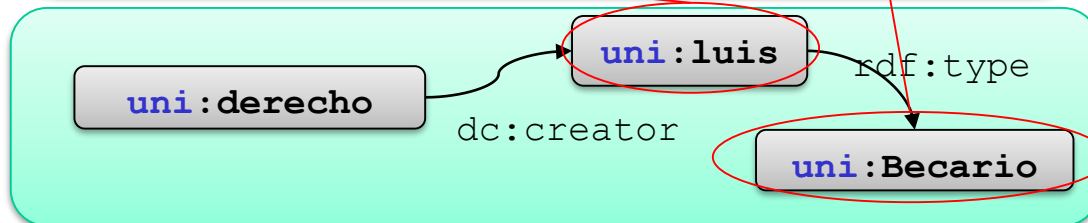
Grafo 1



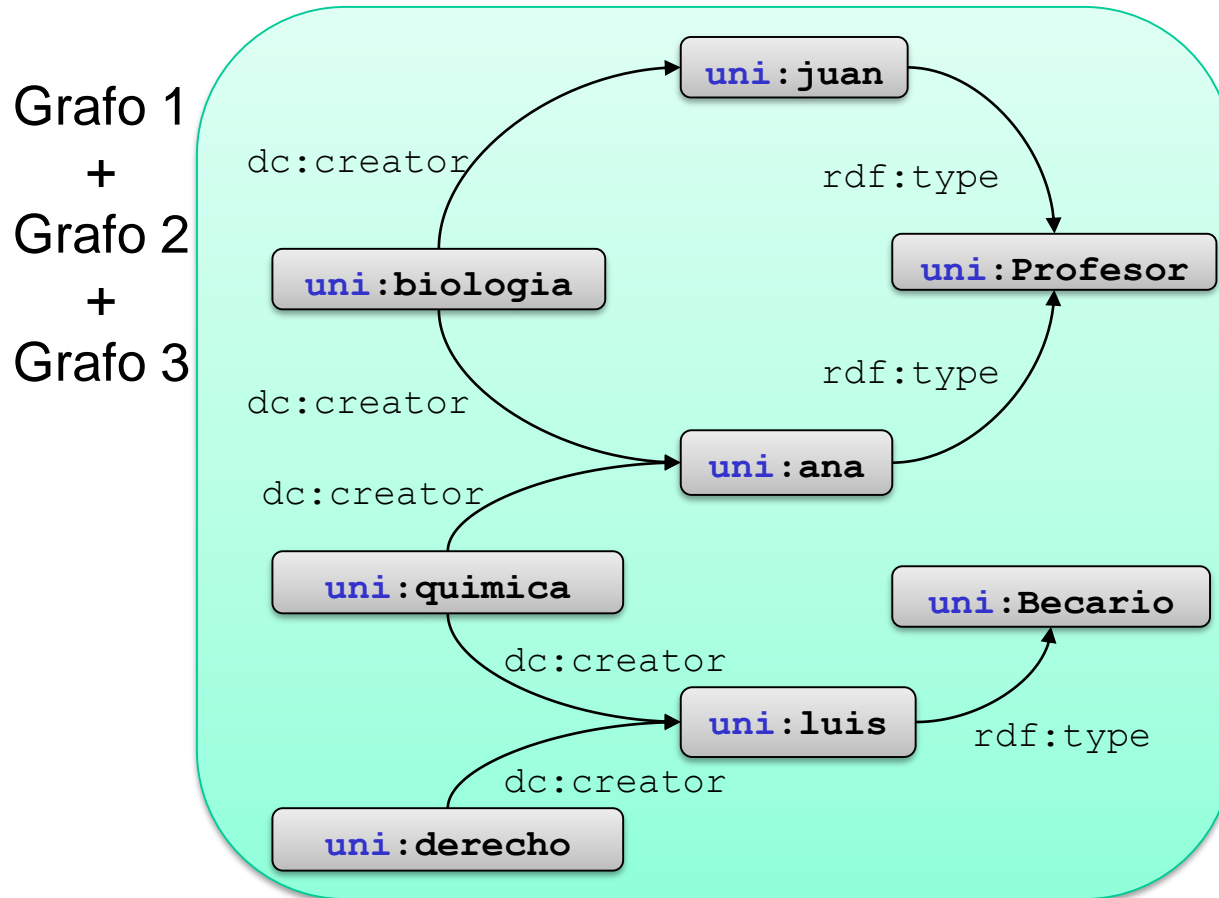
Grafo 2



Grafo 3



RDF es composicional



Formatos RDF



Numerosos formatos y sintaxis:

N3

RDF/XML

N-Triples

Turtle

json-ld

RDFa

etc.

...pero...

¡Lo más importante es el modelo de grafo!

SPARQL



Simple Protocol and RDF Query Language

Lenguaje de consultas para la web semántica

Encaje de grafos

Extrae información de modelos RDF

Un protocolo

Define un mecanismo para invocar un servicio

También define un vocabulario para resultados

SPARQL



Ejemplo:

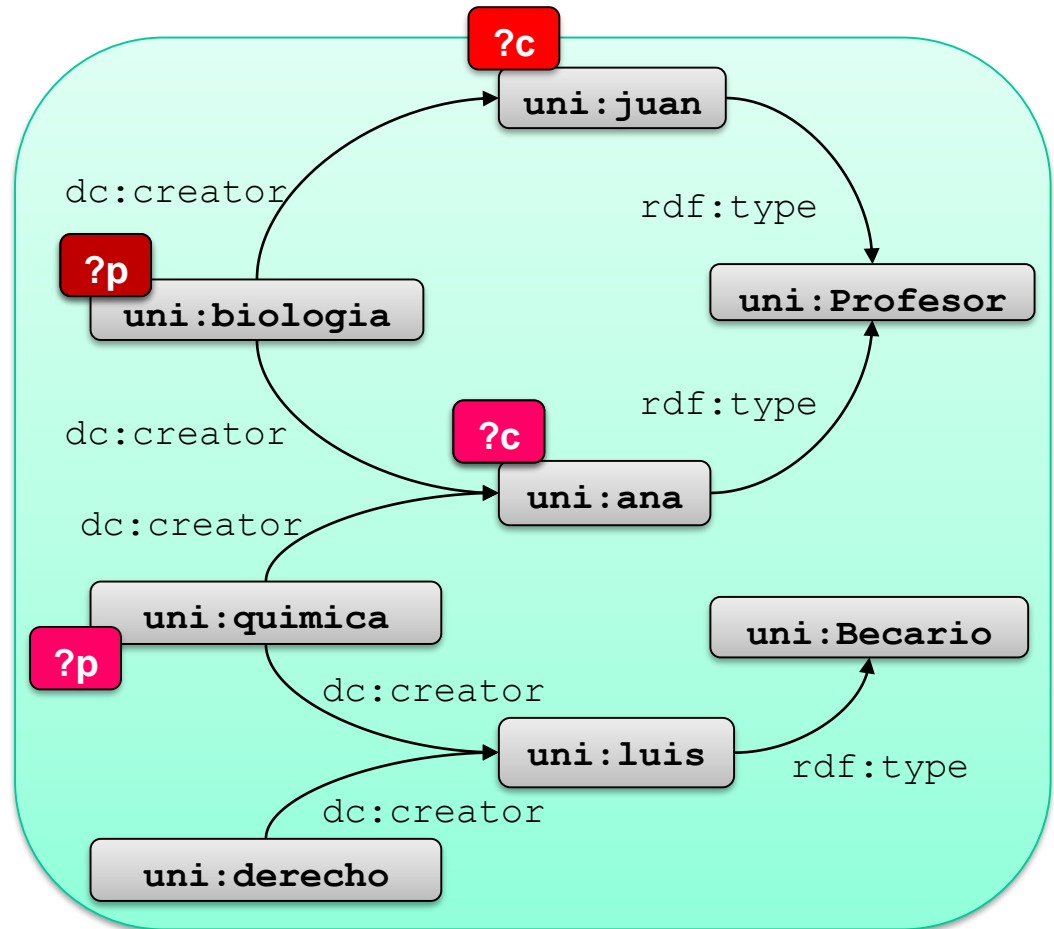
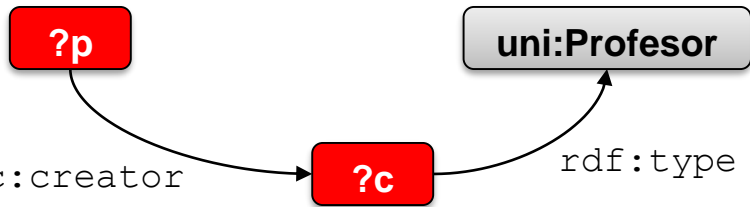
Buscar páginas cuyo autor sea un profesor

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix uni: <http://uniovi.es/>
prefix dc:  <http://purl.org/dc/elements/1.1/>

SELECT ?p ?c WHERE {
  ?p dc:creator ?c .
  ?c rdf:type    uni:Profesor.
}
```

Encaje de grafos

```
SELECT ?p ?c WHERE {  
  ?p dc:creator ?c .  
  ?c rdf:type uni:Profesor  
}
```



Resultados

?p	?c
uni:biologia	uni:juan
uni:biologia	uni:ana
uni:quimica	uni:ana

RDF Schema



Añade un vocabulario de esquema a RDF

Class, Property, Resource,...

type, subclassOf, subPropertyOf,...

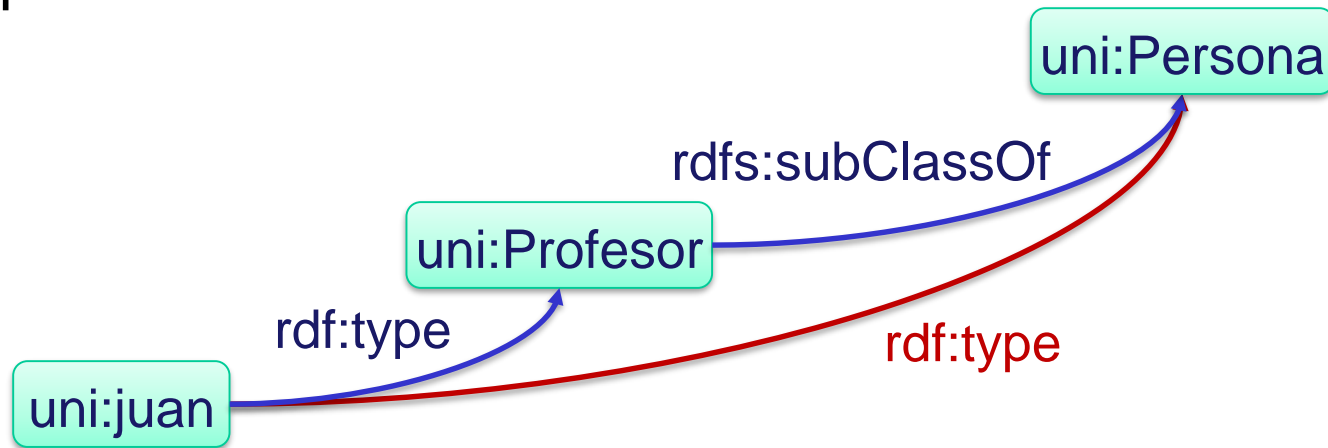
range, domain,...

RDF Schema permite **inferencias**

RDF Schema



Ejemplo



SPARQL + Inferencia



Combinar SPARQL e inferencia

Ejemplo:

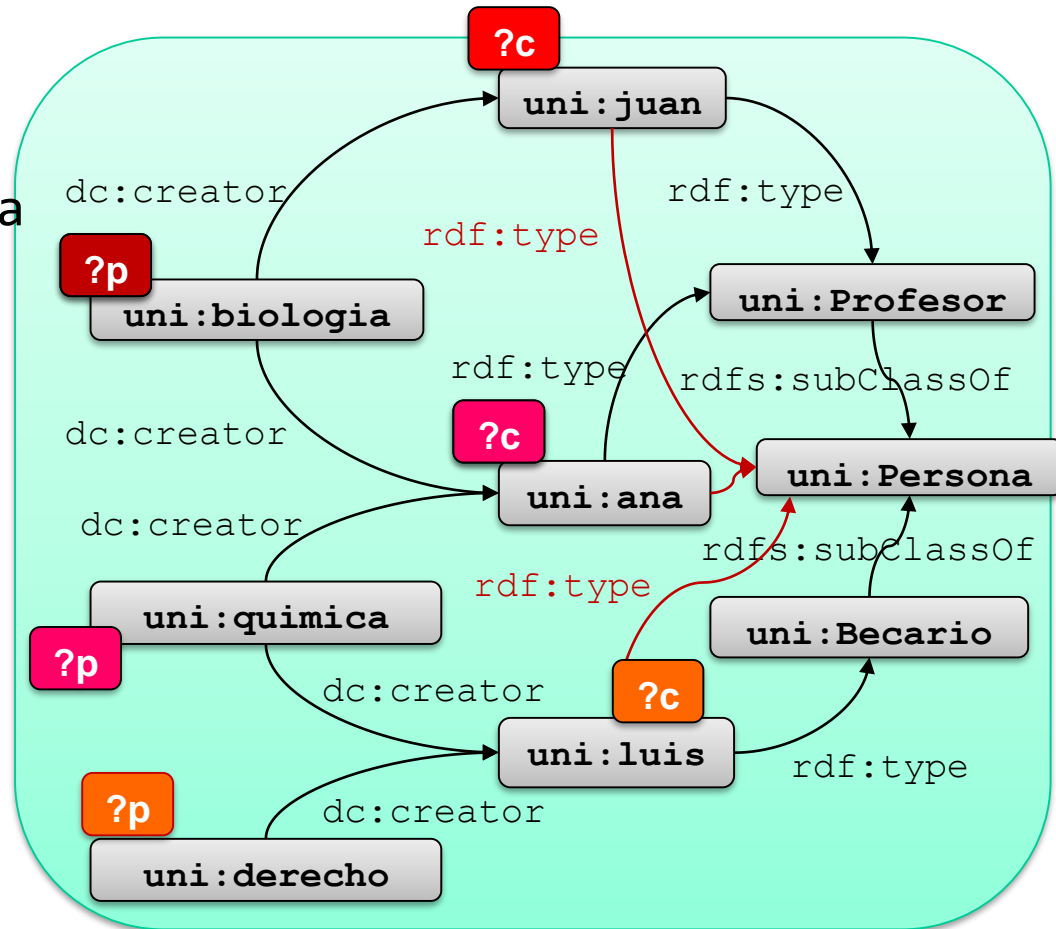
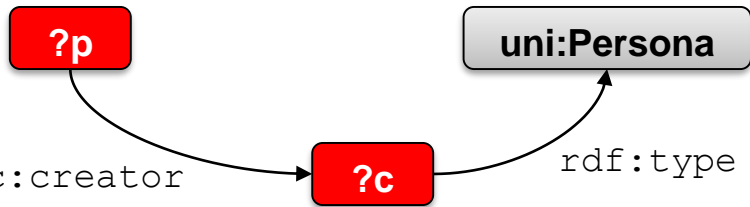
Páginas cuyo autor sea una persona

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix uni: <http://uniovi.es/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

SELECT ?p ?c WHERE {
  ?p dc:creator ?c .
  ?c rdf:type uni:Persona.
}
```

SPARQL con inferencia

```
SELECT ?p ?c WHERE {
  ?p dc:creator ?c .
  ?c rdf:type uni:Persona
}
```



Resultados

?p	?c
uni:biologia	uni:juan
uni:biologia	uni:ana
uni:quimica	uni:ana
uni:derecho	uni:luis



Ontologías

RDF Schema permite hacer inferencias sencillas

Poca expresividad

OWL (Web Ontology Language)

Añade más expresividad

Formalizar dominios concretos: ontologías

Expresividad vs Complejidad



Mitos de la Web Semántica

Navegador inteligente

Una nueva Web

El cerebro global

La gran verdad: Una única ontología

Una etiqueta para cada cosa

Nadie querrá compartir datos

Demasiada apertura

Moda pasajera

No hay *Killer application*



El navegador inteligente

Mito:

El objetivo es conseguir sistemas que naveguen por internet de forma inteligente

Realidad:

Objetivo = desarrollar tecnologías que faciliten el procesamiento automático de la información de la Web y su integración

No es Inteligencia Artificial pero sí se utilizan técnicas de esa disciplina

Una nueva Web

Mito:

La Web Semántica (\approx Web 3.0) es una nueva versión de la web que obligará a cambiar todo lo que ya hay

Realidad:

Se propone transición gradual.

Las tecnologías ofrecerán valor añadido.

El cerebro global

Mito:

El proyecto de la Web semántica generará un cerebro global

Realidad:

La web semántica facilitará un mejor uso de los datos de la web.

Sí es un camino hacia la inteligencia colectiva

La gran verdad

Mito:

Se propone la creación de **una única ontología** con todo el conocimiento de la humanidad

Realidad:

Múltiples ontologías para diferentes dominios

Facilitar la integración

Mejorar la descripción de dominios

Una etiqueta para cada cosa

Mito:

El objetivo es asignar una etiqueta similar a RFID para cada cosa

Realidad

No es factible que cada cosa conlleve sus propios metadatos
Descripciones de recursos externas a ellos

Nadie querrá compartir datos

Mito:

Los proveedores de información no tendrán motivación para adoptar tecnologías nuevas

Realidad:

Lo harán cuando encuentren un retorno de inversión adecuado
Posicionamiento semántico

<http://schema.org>

Principales buscadores indexan datos estructurados
Google, Yandex, Yahoo, Bing

Demasiada apertura

Mito:

Si abrimos datos de bases de datos, los perdemos

Realidad:

Hay tecnologías para limitar acceso

Declarar de dónde provienen los datos

Establecer propiedad legal de los datos

Moda pasajera

Mito:

Mito1: La Web semántica es algo nuevo

Mito 2: La Web semántica es algo viejo

Realidad:

Planteada ya en 1994, visión a largo plazo

Exceso de entusiasmo vs escepticismo

Casos de éxito: RSS, microformatos, XBRL,...

"A little semantics goes a long way"

No hay *killer application*

Mito:

No se ha desarrollado una *killer application*

Realidad:

¿Es necesaria?

¿*Linked Open Data*?

Retos

Proyecto Web semántica:
Primera fase = producción



Segunda fase = consumo



Calidad es cada vez más importante



Fin de la Presentación

