TALENTO E INNOVACIÓN
Escuela Europea de Negocios

Centro colaborador:
UEMC
Universidad Europea Miguel de Cervantes

Web semántica: fundamentos, tecnologías y futuro

Javier Iglesia Aparicio

16 mayo 2018

CENTRO INTERNACIONAL DE TECNOLOGÍAS AVANZADAS PARA EL MEDIO RURAL
CITA

Índice

La información en la web actual

Problemas en la recuperación de información

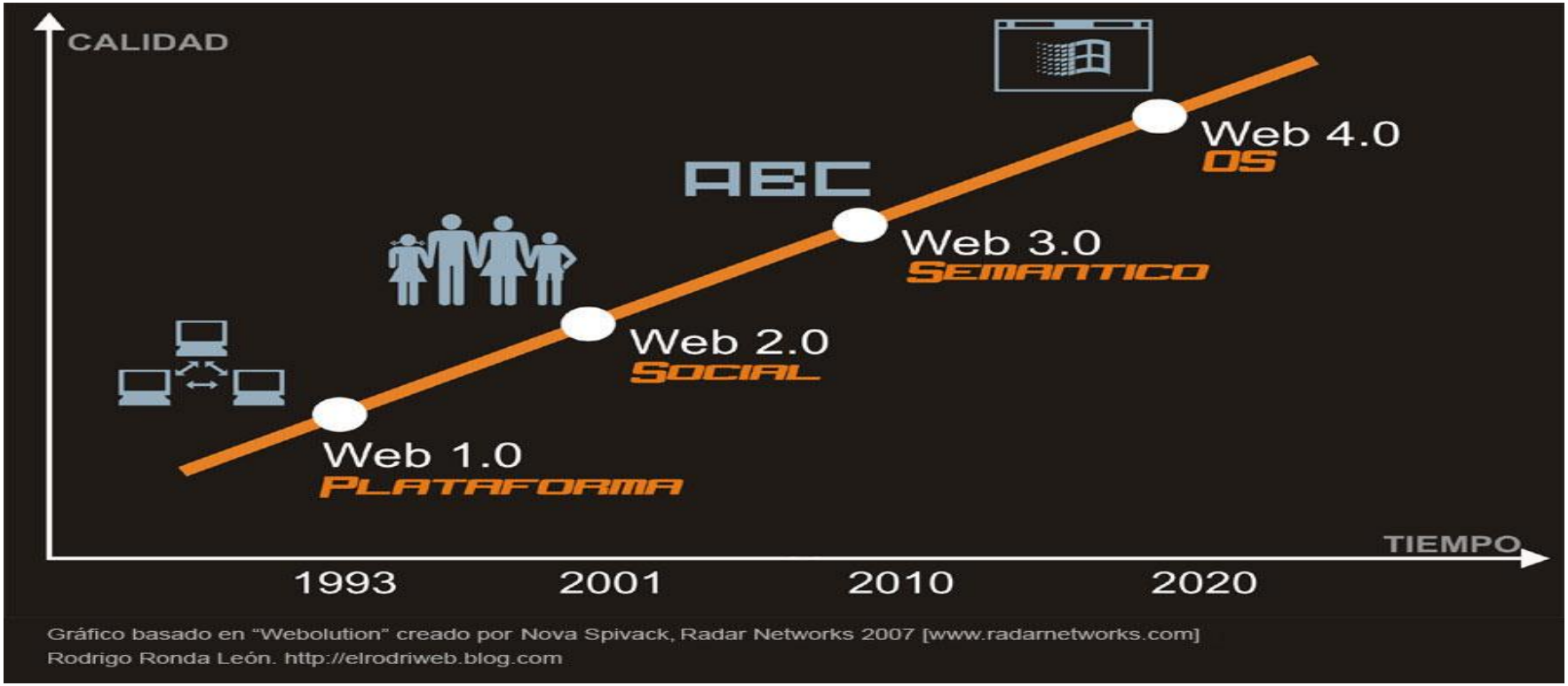
La Web Semántica

Repaso a las tecnologías de web semántica más importantes

Futuro de la web semántica

El “caos” de los datos y la información en la Web

La evolución de la web 1991-2018



Desde 2012: la Web de los datos

Tecnología de web semántica + Big Data + Visualización + Ubicuidad + Concepto Open+ Internet de las Cosas

Algoritmos + Machine Learning

Hacia una web inteligente

Algunos datos enormes

Datos de junio 2017

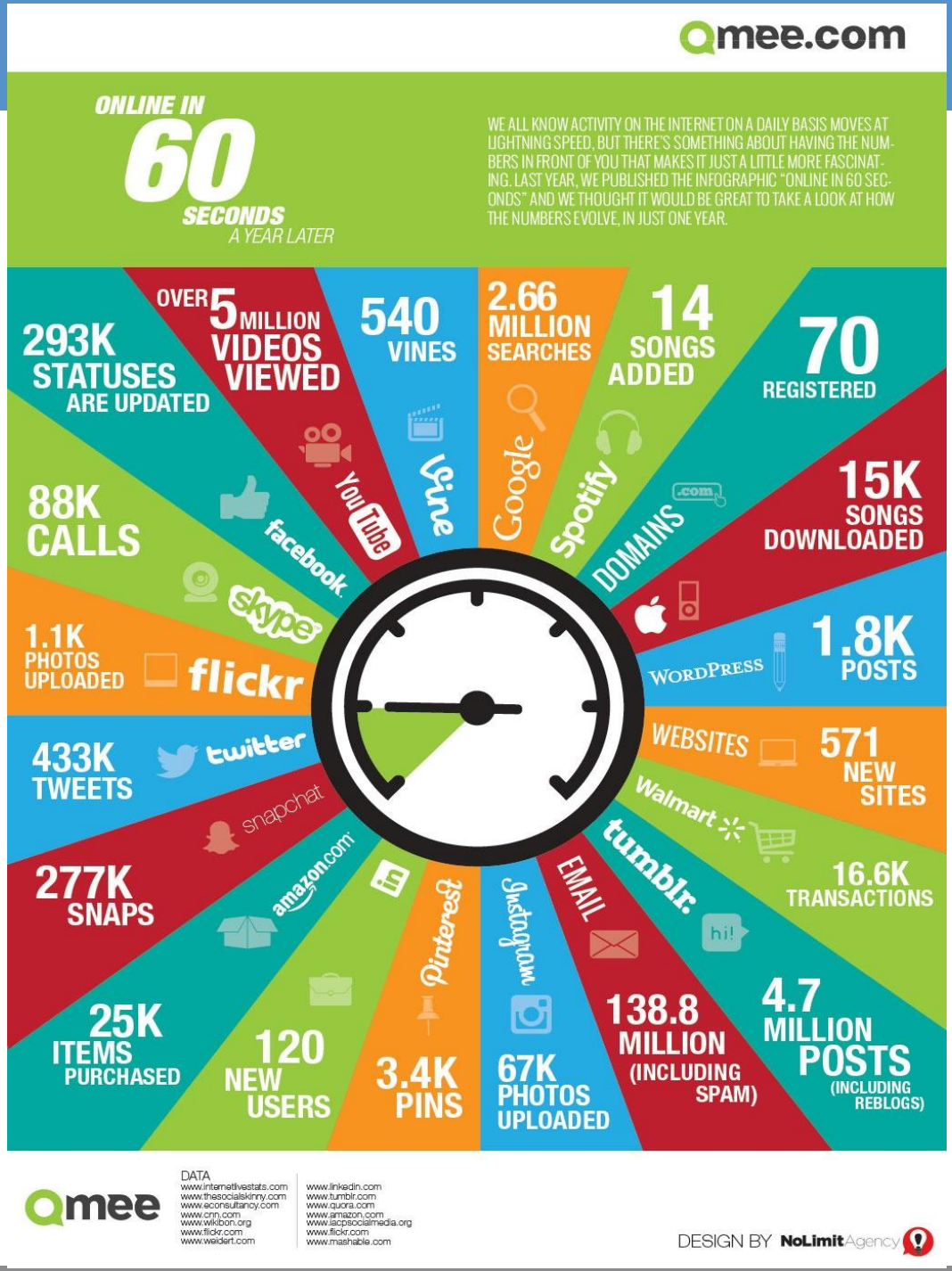
- 3.885 millones de usuarios conectados (360 en el año 2000)

- 30 **ZB** de información (25 filas de libros entre la Tierra y Plutón)

- 1.256 millones de servidores

Fuentes:
http://www.internetlivestats.com/total-number-of-websites/
http://www.internetworldstats.com/stats.htm

En continuo y vertiginoso crecimiento



Internet de las Cosas

El **Internet de las cosas** (Internet of Things en inglés, y abreviado "IoT") consiste en conectar cualquier dispositivo a Internet para que interactúe con una aplicación, con otro objeto, o con nosotros mismos y así obtener datos que podamos utilizar para nuestro beneficio.

Big Data

Una de las consecuencias de la llegada del IoT es la generación masiva de datos: **Big Data**.

Análisis masivo de datos que no pueden ser procesados o analizados utilizando herramientas tradicionales (es decir, que generalmente tienen que ser gestionados por una plataforma específica para ellos).

Se necesitan nuevas herramientas que:

- Permitan almacenar y recuperar grandes volúmenes de datos (Apache SOLR, MongoDB, Hadoop)
- **Procesen los datos originales y los formateen para poder ser analizados**
- Herramientas que permitan crear visualizaciones y análisis a partir de esos datos procesados - Business Intelligence

La necesidad de publicar DATOS ESTRUCTURADOS

Existe algo fundamental para poder facilitar la publicación, la recolección y el procesado de un dato: **El dato debe estar correctamente publicado**.

Puedo leer una tabla donde ponga *Temperatura* y luego varias filas con mediciones. Pero, en qué sistema de medida se encuentran ¿Celsius?¿Fahrenheit? ¿Están las medidas correctamente validadas? ¿Qué nivel de sensibilidad tiene el termómetro? ¿Sabemos si por ejemplo a partir de cierta temperatura no mide correctamente?

Toda esta información se ha de proporcionar junto a los datos. Son los **metadatos**, la información necesaria para poder entender, analizar y comprender correctamente los datos brutos y generar así información.

Dato + Metadatos = Dato estructurado

Objetivo final: una Web Inteligente

- Que disponga de multitud de datos abiertos
- Que esos datos sean estructurados: proporcionen suficiente información sobre sí mismos para poder realizar operaciones y utilizar la lógica
- Que existan nodos especializados en cada tipo de datos
- Que los distintos nodos estén enlazados entre sí de modo semántico
- Que existan aplicaciones capaces de entender todas esas descripciones de datos, consultarlas e inferir respuestas.

Problemas de la búsqueda de información en la Web actual


CLAUDE
PRESBYTERIAN CHURCH
THERE ARE SOME QUESTIONS THAT CAN'T BE ANSWERED BY GOOGLE
Citytv
CityNews

Problemas de la presentación de la información en la web actual

- HTML no presenta datos estructurados, solo formatean los contenidos y los datos
- No tiene mecanismos de procesado automático de la información
- No hay mecanismos de interoperabilidad completa de los sistemas de información:
    - **Interoperabilidad sintáctica**: los datos no están estructurados de acuerdo a un formato de entiendan todas las máquinas
    - **Interoperabilidad semántica**: no es posible trabajar con el significado de los textos. Las máquinas no entienden de sinónimos, homónimos, polisemias…

¿Cómo funciona un buscador actualmente?

1. Una serie de robots o arañas web rastrean las webs

2. Las palabras extraídas se indexan y se asigna una relevancia a los sitios de donde son extraídas según multitud de parámetros:

   a. **Frecuencia de aparición de una o varias palabras en un texto**
   b. Aparición de palabras relacionadas
   c. Posición dentro de documento
   d. Los enlaces que el documento tiene con otros web
   e. Los enlaces externos que llevan a la web
   f. Antigüedad y autoridad del dominio

3. Una interfaz gráfica permite la consulta de estos datos y se presentan resultados

¿Cómo da los resultados un buscador actualmente?

Usted realiza una búsqueda...

1 Se envía al servidor web de Google...

El servidor de Google está formada por una red de más de 10.000 equipos conectados a Internet.

2 La petición se envía a los servidores de índices.

Los servidores de íncide indican en que servidores de datos se encuentran las webs que contienen las palabras de la petición.

3 La petición viaja hasta los servidores de documentos que contienen una copia de todas las webs que ha visitado Google.

Las páginas se ordenan aplicando el algoritmo de PageRank que calcula la imprtancia de la web en función de más 500 millones de variables.

4 Con la información obtenida se genera la página de resultados de la búsqueda.

5 Finalmente se devuelve el resultado de la búsqueda.

Todo el proceso se completa en menos de un segundo.

Resultado: búsquedas no siempre precisas

- Poco precisas, porque en ocasiones no somos demasiado explícitos en nuestras búsquedas
- Muy dependientes del vocabulario empleado en la búsqueda, del idioma, etc.
- Obvian todos aquellos documentos que no tienen texto o metadatos rastreables: audio, texto dentro de un vídeo o de un flash…
- Los resultados suelen ser pésimos para consultas muy específicas y con muchos condicionantes

¡¡Queremos resultados como este!!

¿Solución?

Cambiar el modo de presentación de los datos en las páginas web.

Hay que proporcionar más datos que ayuden a las máquinas (a los buscadores) a procesar la información para que nos den mejores respuestas

Ejemplo

Página 1

… El canguro cuida del niño …

Página 2

… El Niño afecta a los canguros…

Ejemplo - Sin web semántica (HTML)

El canguro cuida del niño

El Niño afecta a los canguros

Página 1: El canguro cuida del niño

Página 2: El Niño afecta a los canguros

El rastreador web ve lo siguiente:

	Canguro	Niño
Página 1	1	1
Página 2	1	1

Resultado:

Ambas páginas hablan de lo mismo

Ejemplo - Con web semántica

El canguro cuida del niño
El Niño afecta a los canguros

Página 1:

El <profesion>canguro</profesion>
cuida del <persona>niño</persona>

Página 2:

El <clima clase="fenomeno">
Niño</clima> afecta a los
<animal class="marsupial"> canguros
</animal>

El rastreador web ve lo siguiente:

	Personas		Animales	Clima
	Profesión	Edad	Marsupiales	Fenómeno atmosférico
	canguro	niño	canguro	Niño
Página 1	1	1		
Página 2			1	1

Resultado:

Cada página habla de distintos conceptos

Ejemplo - Comparación

El canguro cuida del niño
El Niño afecta a los canguros

HTML	WEB SEMÁNTICA
2 palabras	4 conceptos

La Web Semántica y sus tecnologías

Un primer intento: los microformatos

Un microformato es una forma simple de agregar significado semántico a un contenido legible por el humano y que para la máquina es sólo texto plano.

hCard

```
<div class="vcard">
    <div class="fn">Juan Pérez</div>
    <div class="org">El Ejemplo S. A.</div>
    <div class="tel">604-555-1234</div>
</div>
```

Mantenidos y creados por la comunidad de http://microformats.org/

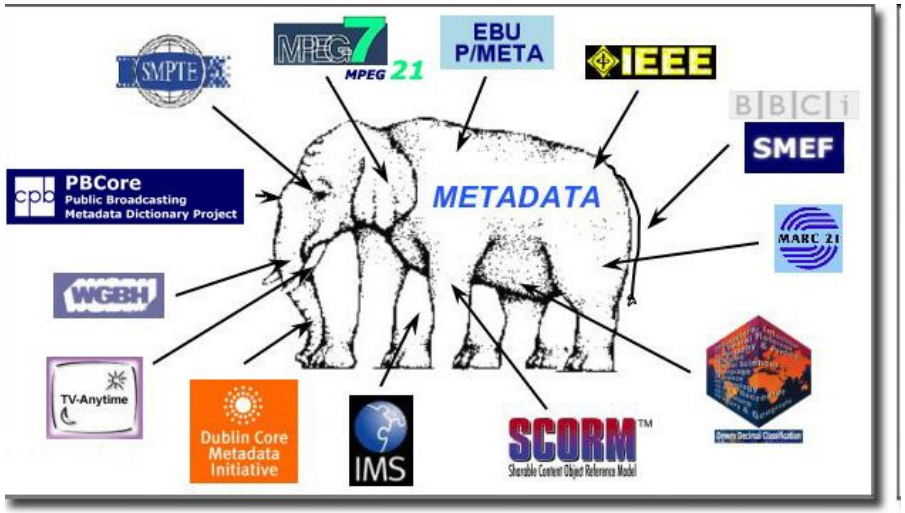
Un avance más ambicioso: los metadatos...

Datos sobre los datos basados en estándares

Suelen ser muy específicos

Mucha diversidad y sin compatibilidad entre sí

Se pueden combinar con tesauros o vocabularios controlados



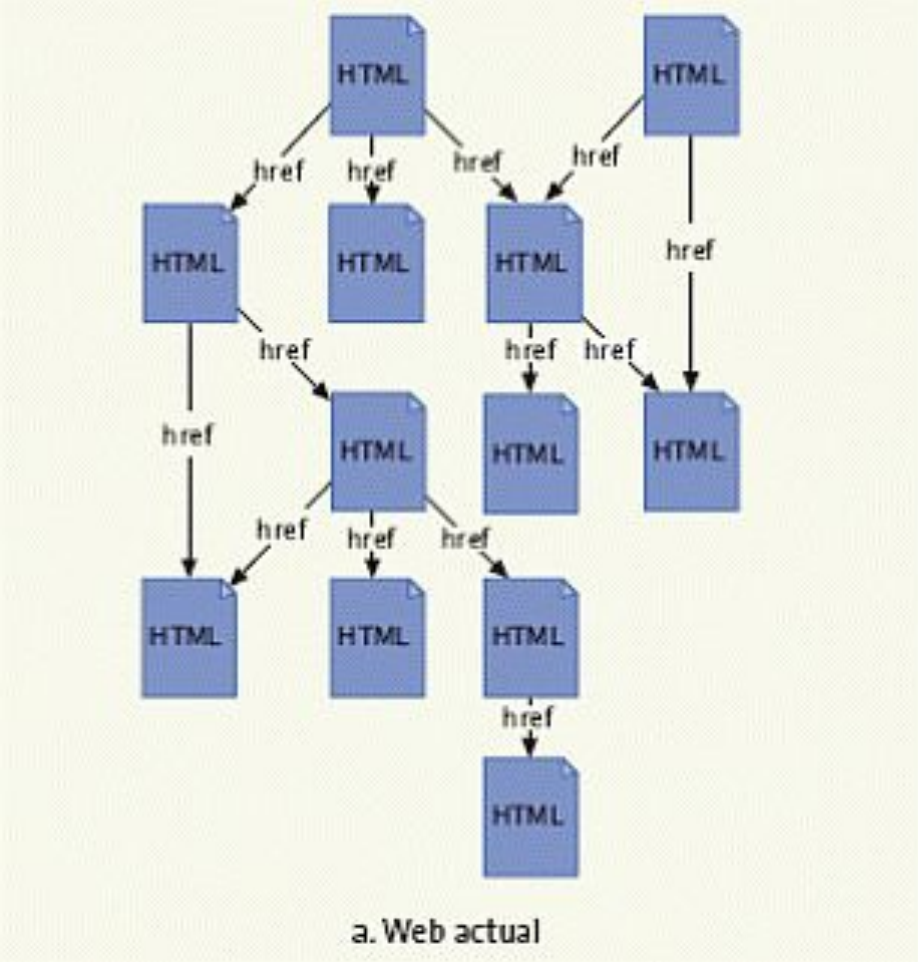
El intento de verdad: La Web semántica

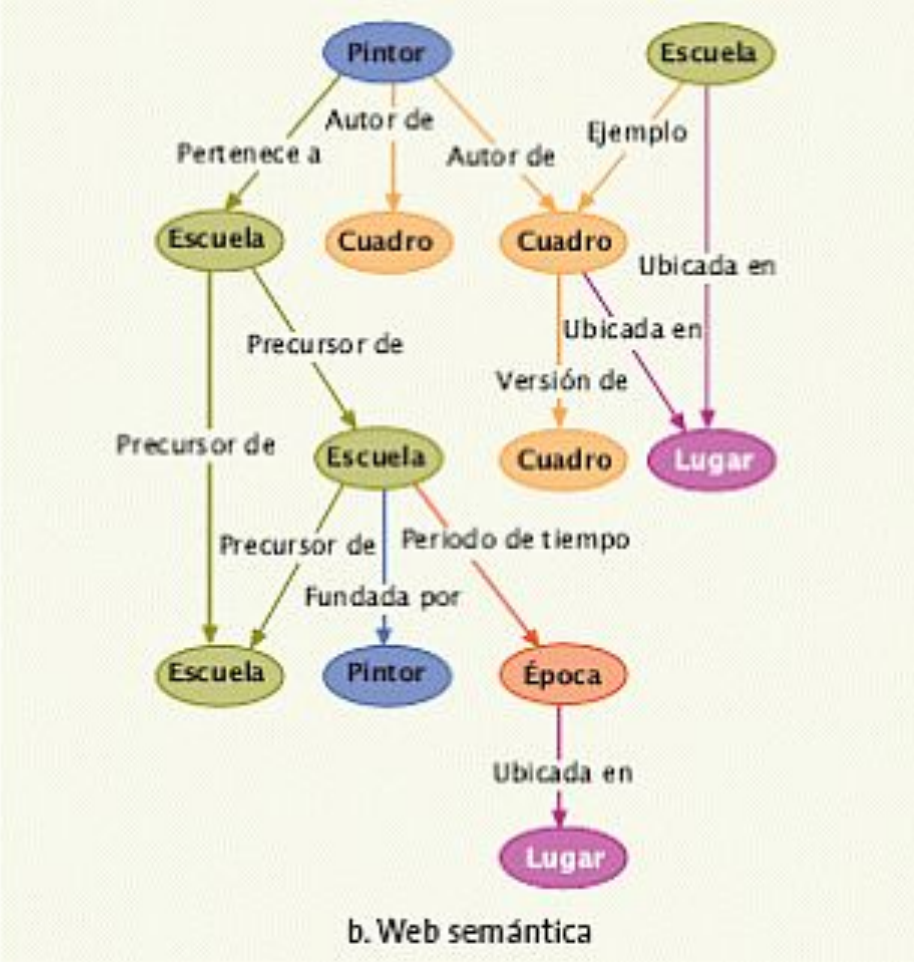
Se basa en añadir metadatos semánticos y de ontologías a las páginas web para describir el contenido, el significado y la relación de los datos

Es la **web de los datos**. Más concretamente, la web de los datos relacionados.

El objetivo es crear un medio universal que permita el intercambio de datos y brindar un mayor significado a la misma para que puedan ser interpretadas por las máquinas.

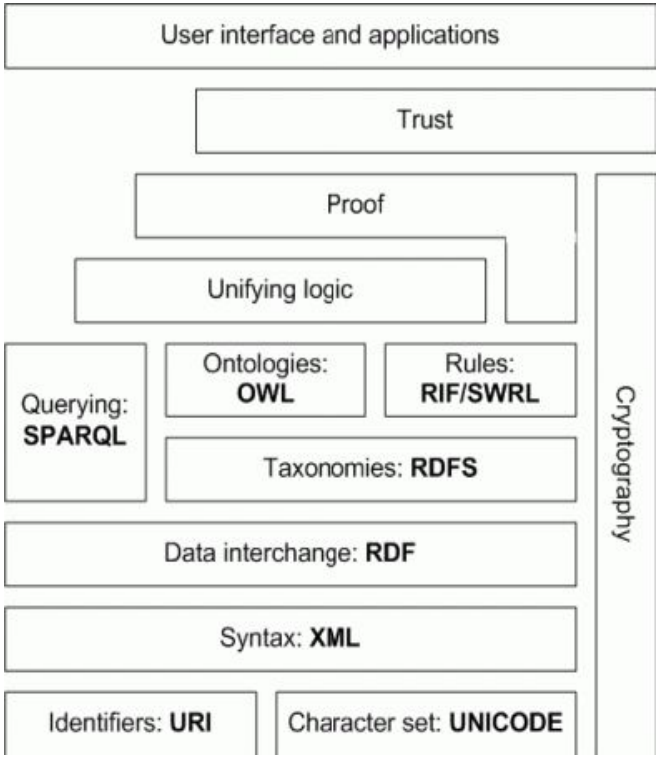
El intento de verdad: La Web semántica

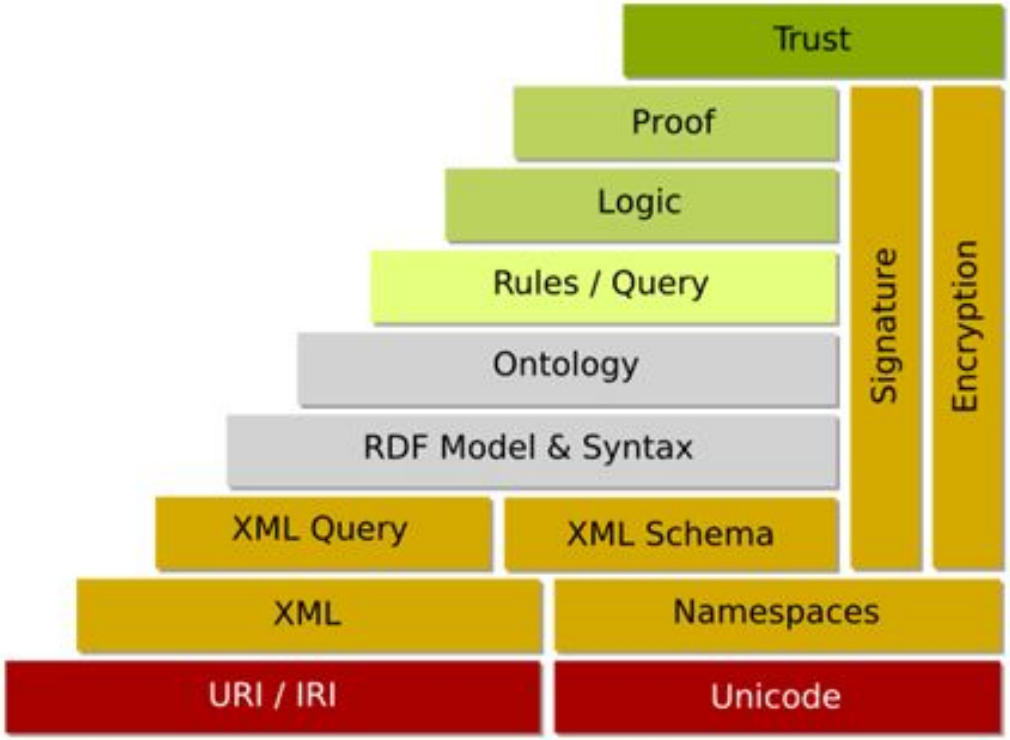




Estándares de la Web Semántica

Un formato universal para describir recursos de cualquier tipo: RDF
Un formato para describir ámbitos de conocimiento (ontologías): SKOS, OWL
Lenguajes para preguntar por los datos: SPARQL
Lógica para que las máquinas puedan hacer inferencias





El concepto de ontología

La ontología es una representación formal y explícita de la estructura conceptual del campo sobre el que se trabaja.

La idea es que la Web semántica está formada, al menos en parte, por una red de nodos tipificados e interconectados mediante clases y relaciones definidas por una ontología compartida por sus distintos autores.

- **Son muy complejas de describir, necesitan de especialistas en el ámbito del conocimiento y de consenso.**
- **Profundo conocimiento de la lógica**

OWL: El lenguaje para describir ontologías

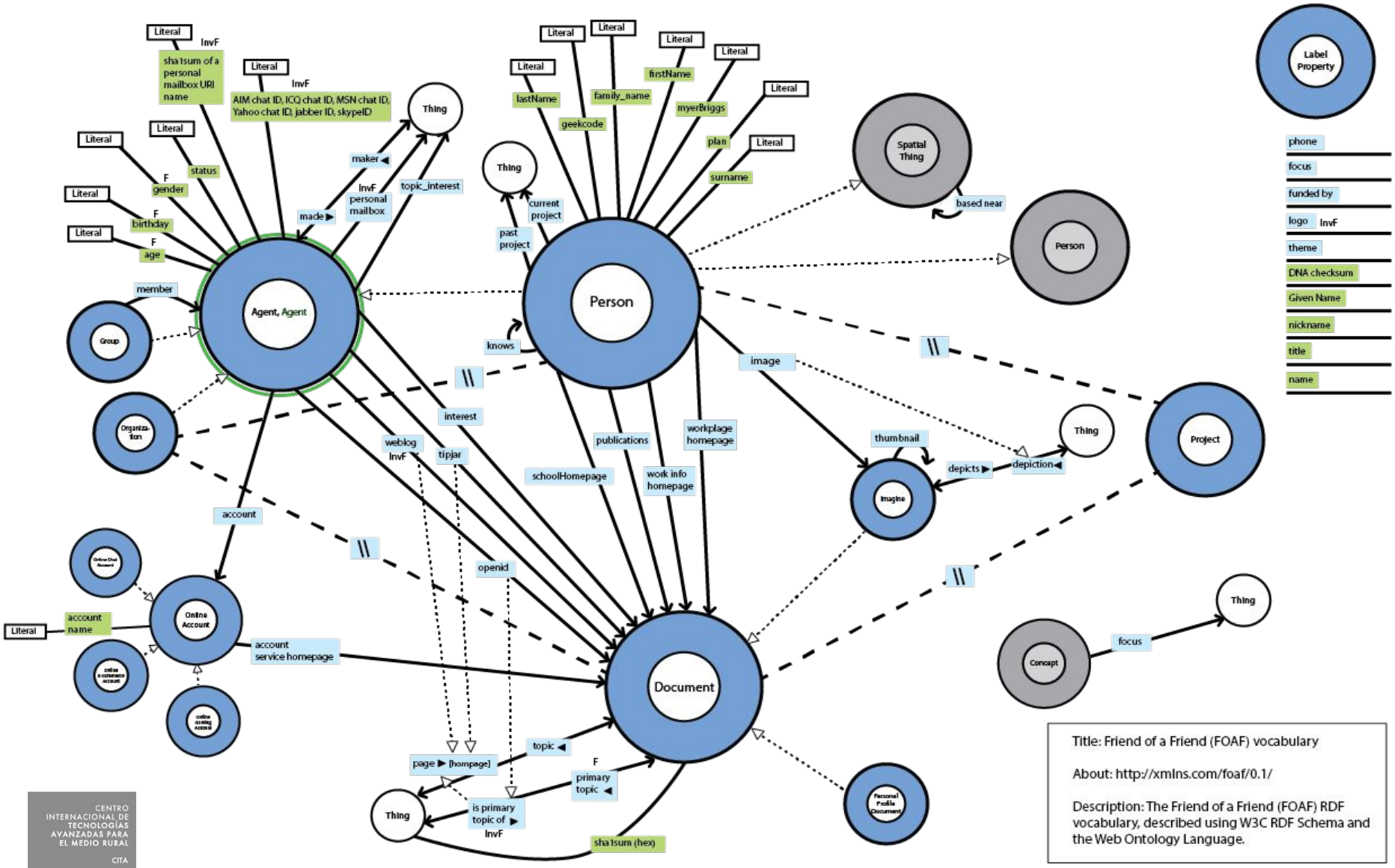
OWL (*Web Ontology Language*) es un mecanismo para describir temas o vocabularios específicos.

Lenguaje para definir ontologías estructuradas que pueden ser utilizadas a través de diferentes sistemas. OWL define clases, propiedades e individuos y las relaciones entre esos elementos.

Con OWL se puede definir que “un libro de papel es distinto de un libro electrónico” o que un libro “puede tener uno o más autores” y que “un autor puede haber escrito uno o más libros”.

De esta forma se podrá “preguntar” a la ontología y se podrá razonar sobre ella para obtener respuestas adecuadas.

Veamos una ontología por dentro: FOAF (http://visualdataweb.de/webvowl/)



RDF (Resource Description Framework): un lenguaje para describir cualquier tipo de documento

RDF permite describir recursos de una forma similar a los metadatos. Los recursos pueden ser páginas web pero, también otras cosas como valores estáticos definidos en otros estándares de metadatos como, por ejemplo, Dublin Core.

Una descripción RDF es una tripleta, es decir, tiene tres partes:

- Lo que se describe, el recurso, es el **Sujeto**
- Las propiedades de lo que se describe o **Predicado**
- Los valores de las propiedades de lo que se describe u **Objeto**

De esta forma se van describiendo **grafos** de cada recurso.

<sujeto> tiene una propiedad <predicado> cuyo valor es <objeto>

RDF: un lenguaje para describir cualquier tipo de documento

Veamos cómo funciona con un sencillo ejemplo: El libro *La Divina Comedia* fue escrito por *Dante*. Identifiquemos las tres partes esenciales de esta descripción:

- El libro *La Divina Comedia* es lo que estoy describiendo. Es el sujeto.
- El libro *La Divina Comedia* tiene una propiedad: autor. Es el predicado.
- La propiedad autor tiene un valor: *Dante Alighieri*. Es el objeto.

Y a partir de estos datos puedo inferir más cosas: lo que describo es un Libro y un Libro tiene una propiedad llamada título.

RDF: un lenguaje para describir cualquier tipo de documento

Si ahora tenemos enlazado un documento PDF en la Web que contiene el texto de *La Divina Comedia* podemos describirlo mediante RDF haciendo uso de los metadatos de Dublin Core.

Esta sería la sintaxis RDF:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc=http://purl.org/dc/elements/1.0/ >

    <rdf:Description
    rdf:about="http://www.librosgratisweb.com/pdf/alighieri-dante/divina-comedia.pdf">

        <dc:title>La Divina Comedia</dc:title>
        <dc:creator>Dante Alighieri</dc:creator>

    </rdf:Description>
</rdf:RDF>
```

SPARQL

SPARQL (*Simple Protocol and RDF Query Language*) es un lenguaje de consulta sobre RDF, que permite hacer búsquedas sobre los recursos de la Web Semántica utilizando distintas fuentes datos.

Sintaxis básica

PREFIX (Namespace Prefixes)

```
e.g. PREFIX plant: <http://www.linkeddatatools.com/plants>
```

SELECT (Result Set)

```
e.g. SELECT ?name
```

FROM (Data Set)

```
e.g. FROM <http://www.linkeddatatools.com/plantsdata/plants.rdf>
```

WHERE (Query Triple Pattern)

```
e.g. WHERE { ?planttype plant:planttype ?name }
```

ORDER BY, DISTINCT etc (Modifiers)

```
e.g. ORDER BY ?name
```

El concepto de Dato Enlazado

Todos conocemos el concepto de enlace, básico en la web. Pero esta relación no aporta ningún tipo de información a las máquinas de por qué se realiza esa relación.

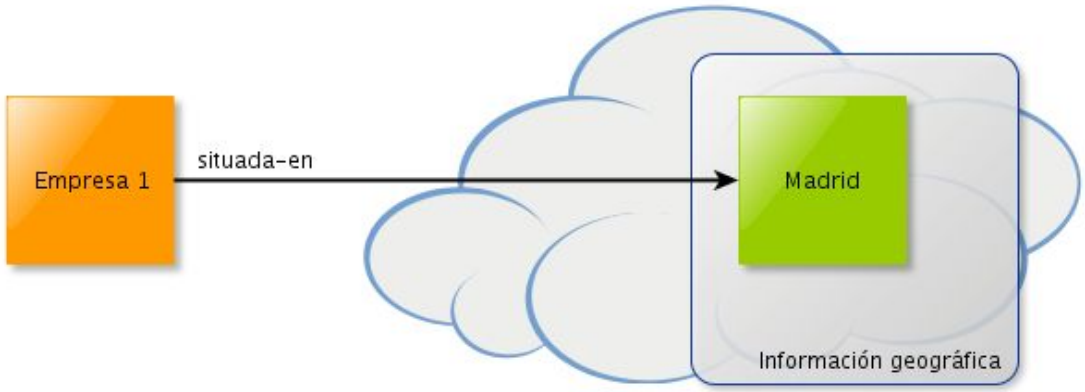
Es necesario que el enlace, a nivel de código, aporte información adicional que la explique.

HTML

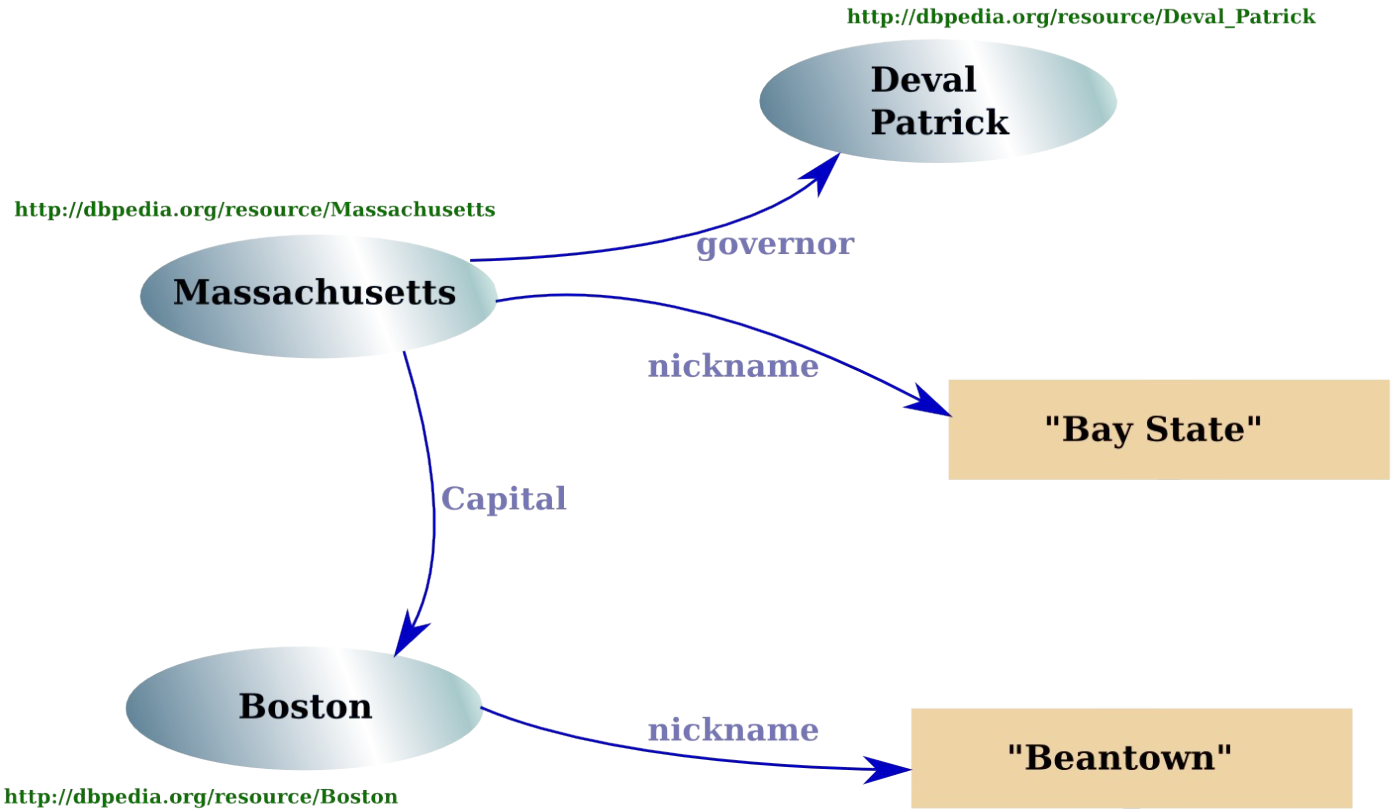
Visita las oficinas de ACME en Madrid

Dato Enlazado

Visita las oficinas de <empresa name="ACME">ACME</empresa> en <empresa situada-en="Madrid" type="Geonames">Madrid</empresa>



Ejemplo de representación de un dato enlazado



Ejemplo de representación de un dato enlazado

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:db="http://dbpedia.org/resource/">
  <rdf:Description rdf:about="http://dbpedia.org/resource/Massachusetts">
    <db:Governor>
      <rdf:Description rdf:about="http://dbpedia.org/resource/Deval_Patrick" />
    </db:Governor>
    <db:Nickname>Bay State</db:Nickname>
    <db:Capital>
      <rdf:Description rdf:about="http://dbpedia.org/resource/Boston">
        <db:Nickname>Beantown</db:Nickname>
      </rdf:Description>
    </db:Capital>
  </rdf:Description>
</rdf:RDF>
```

Estado actual y futuro de la Web Semántica

¿Qué permitirá la Web Semántica?

Obtener resultados y respuestas más precisas a nuestras consultas

"Búscame todos los electricistas existentes a 5 kilómetros a la redonda de mi posición y lístame primero los más baratos"

- ***Buscadores semánticos***.
- ***Agentes personales inteligentes***.
- ***Aplicaciones de integración de fuentes de datos heterogéneas***.
- ***Aplicaciones de anotación semántica de contenidos multimedia***. Permiten catalogar los contenidos multimedia de forma semántica, pudiéndose realizar catálogos de contenidos personalizados, descubrir nuevos recursos multimedia de interés para el usuario, etc.
- ***Aplicaciones de adaptación automática de contenidos basándose en la anotación semántica de los mismos***. La idea que subyace en estas aplicaciones es que los contenidos web sean adaptados dinámicamente teniendo en cuenta su semántica y la personalización asociada al usuario.

Dificultades para el establecimiento de la web semántica

Para que la Web semántica pueda realizarse es importante que:

- **Guarde, al menos al principio, una compatibilidad con la tecnología actual.** Es deseable, por ejemplo, mantener el lenguaje HTML, u otros lenguajes compatibles con los navegadores actuales, como vehículo de comunicación con el usuario. La asociación entre las instancias de la Web semántica y el código HTML se puede establecer mediante distintos mecanismos. Dos opciones:
  - Conservar los documentos actuales, y crear las instancias asociadas anotando su correspondencia con los documentos. Esta posibilidad es la más viable cuando se parte de un gran volumen de material antiguo.
  - Generar dinámicamente páginas web a partir de las ontologías y sus instancias.
- **La transición de la Web actual a la Web semántica puede implicar un coste altísimo si se tiene en cuenta el volumen de contenidos que ya forman parte de la Web**. Crear y poblar las ontologías supone un esfuerzo extra que puede resultar tedioso cuando se agregan nuevos contenidos, pero directamente prohibitivo en lo que respecta a integrar los miles de contenidos antiguos.
- Los **gestores de contenidos deberán de integrar posibilidades de etiquetado semántico**, bien sea de forma automática o manual, consultando ontologías públicas.
- **Consensuar ontologías en una comunidad**.

Necesitamos...

- **Una tecnología para tratar de convertir lo que se ha escrito hasta ahora.**

- **Vocabularios especializados**, ontologías temáticas, que puedan ser referenciadas desde cualquier contenido

- **Herramientas** que:
  - Permitan catalogar en RDF
  - Permitan enlazar un documentos con entidades: gestores de contenidos semánticos
  - Herramientas que puedan entender todos los datos semánticos y dar respuestas exactas

Esfuerzos de semantización. Schema.org

Schema.org (http://schema.org/). Impulsado por los tres grandes motores de búsqueda: Google, Bing y Yahoo! Search. Su premisa es la siguiente: dado que la expansión de las aplicaciones Web 2.0 ha multiplicado exponencialmente la cantidad de información existente en la web, los motores de búsqueda tienen que introducir nuevos parámetros para calcular la relevancia, detectando duplicaciones, copias, etc.

La solución que proponen es incluir dentro del código HTML información adicional mediante metadatos y microdatos. Para ello crearon un esquema de categorías (no llega a ser una ontología) –que se puede consultar en este enlace, http://schema.org/docs/full.html - para que los webmasters puedan, de alguna manera, clasificar la información de una página web.

Este esquema general no puede modificarse por los usuarios pero éstos sí pueden extenderlo si no encuentran la clasificación adecuada para su contenido. Lo que ofrece Schema.org es una opción de **marcado semántico** de una página web.

En la actualidad ya existen multitud de páginas web y de gestores de contenidos que utilizan schema.org para aportar información adicional.

Ejemplos de schema.org - Escritura Microdata

Existen tres modos de escribir este marcado semántico en la web que vamos a explicar siguiendo el mismo ejemplo. Supongamos que queremos describir que estamos hablando sobre un libro, *El Quijote de la Mancha* de Miguel de Cervantes y aportamos el título, el autor y la fecha de su primera publicación.

```
<div itemscope itemtype="http://schema.org/Book">
<p> Título: <span itemprop="name">El Quijote</span></p>
<p> Autor: <span itemprop="author">Miguel de Cervantes</span></p>
<p> Fecha de publicación: <span itemprop="dateCreated"> 1605
</span></p>
</div>
```

Necesidad de disponer vocabularios específicos publicados

Para poder publicar datos enlazados, además de la existencia de ontologías y de datos abierto es necesario que estos vocabularios estén publicados de forma unívoca en Internet.

Cada nodo se dedicará a un ámbito específico de conocimiento

¿Cómo sabemos cuántos nodos hay?

¿Cómo descubrimos vocabularios ya publicados?

Existen buscadores de vocabularios:

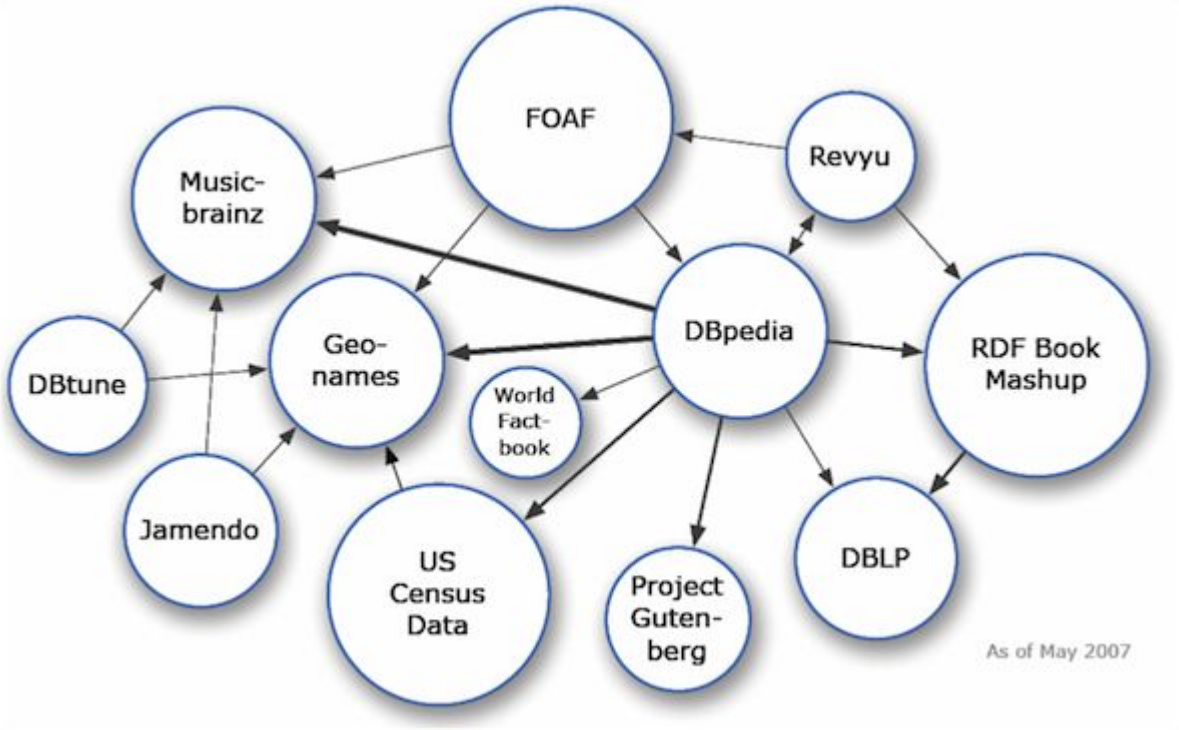
- **LOV (Linked Open Vocabularies)** accesible en http://lov.okfn.org
- **Open Metadata Registry** (http://metadataregistry.org/)

Vocabularios

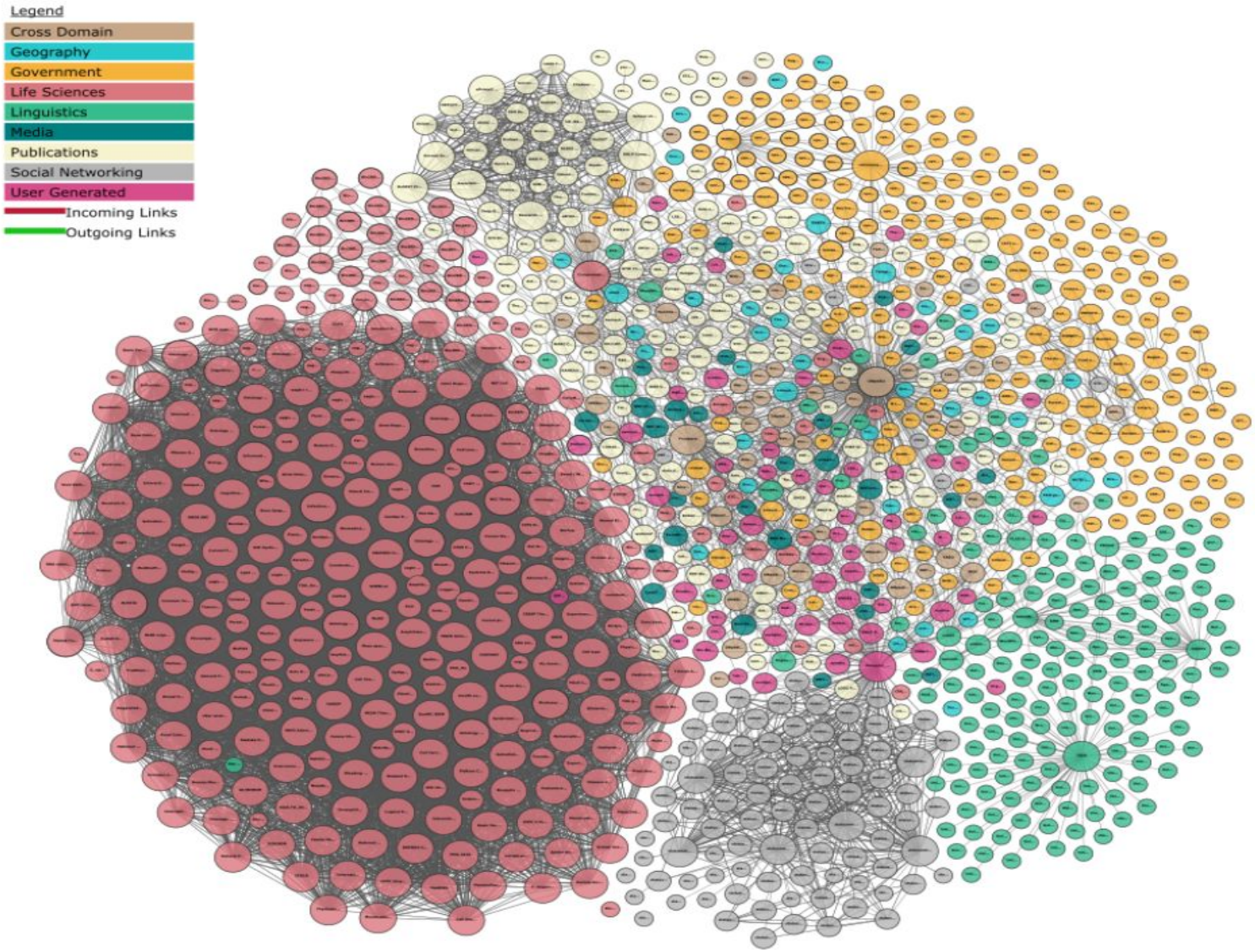
- **VIAF, Fichero de Autoridades Virtual Internacional (https://viaf.org/)**
- **ISNI (http://isni.oclc.org/)**
- **WorldCat**
- **LEMBP Lista de Encabezamientos de Materia**

Y nodos centrales como DBPedia, Geonames, etc...

Los inicios de LOD



La red Linked Open Data (http://lod-cloud.net/)



La semantización del buscador Google

Jupiter - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jupiter**
Jupiter is the fifth planet from the Sun and the largest planet within the Solar System. It is a gas giant with mass one-thousandth that of the Sun but is two and a ...
↳ Mythology - Moons - Galilean moons - Great Red Spot

Images for **júpiter** - Report images

Jupiter - Solar System Exploration: Planets: **Jupiter**: Overview
solarsystem.nasa.gov › Planets
16 May 2012 – General features and data about the planet and its satellites.

El Planeta **Júpiter** y sus satélites
www.xtec.cat/~rmolins1/solar/es/**jupiter**.htm - Translate this page
Júpiter tiene un tenue sistema de anillos, invisible desde la Tierra. También tiene 16 satélites. Cuatro de ellos fueron descubiertos por Galileo en 1610. Era la ...

Júpiter
www.solarviews.com/span/**jupiter**.htm - Translate this page
Características, anillos que posee, composición de su atmósfera, animaciones, imágenes, reseñas y descubridores de las lunas.

Jupiter

en.wikipedia.org

Jupiter is the fifth planet from the Sun and the largest planet within the Solar System. It is a gas giant with mass one-thousandth that of the Sun but is two and a half times the mass of all the other planets in our Solar System combined.
Wikipedia

Distance from Sun: 483,766,802 miles (778,547,200 km)
Gravity: 24.79 m/s²
Density: 1.33 g/cm³
Length of day: 0d 9h 56m
Length of year: 12 years
Mass: 11.34E27 kg (1,899 Earth mass)

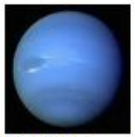
People also search for


Saturn


Mars

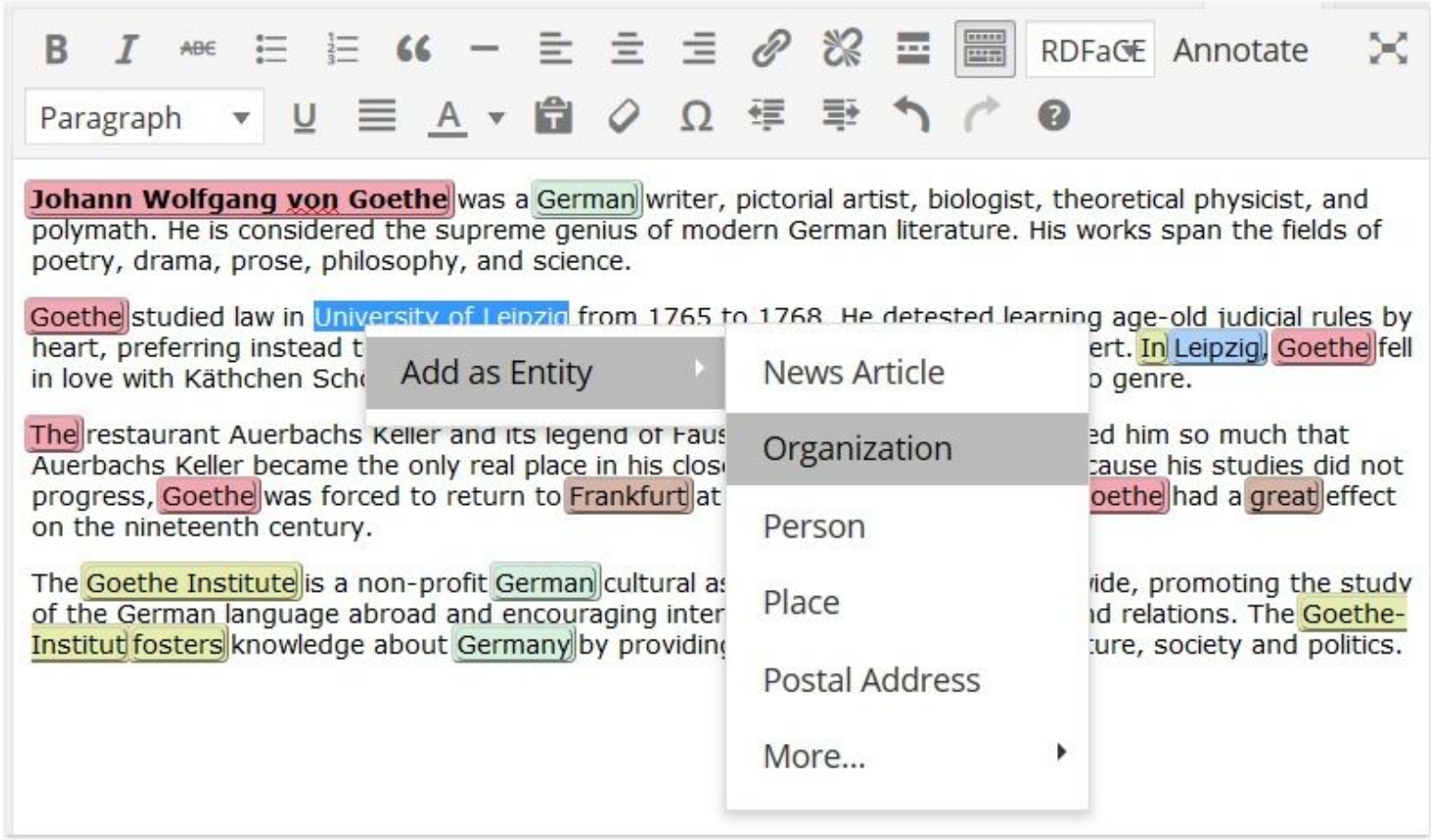

Venus


Uranus


Neptune

Feedback

¿Hay gestores de contenido semánticos?



RDFaCE

GRACIAS

Gracias por su atención
Javier Iglesia