Modelos de información

Ramón Alcarria Garrido

Tomás Robles Valladares

Miguel Ángel Manso Callejo

Borja Bordel Sánchez

GISAI
Grupo de Investigación en Servicios Avanzados de Internet

Universidad Politécnica Madrid
POLITÉCNICA

Introducción a la Internet de las Cosas

Departamento de Ingeniería de Sistemas Telemáticos (UPM)

PROGRAMA

- Big Data
  - MapReduce
- Semántica
  - Modelo de Capas
  - RDF
  - Otras tecnologías

BIG DATA

- Tratamiento y análisis de enormes repositorios de datos donde resulta imposible tratarlos con las herramientas de bases de datos y analíticas convencionales

BIG DATA

- Motivado por la proliferación de páginas web, aplicaciones de imagen y vídeo, redes sociales, dispositivos móviles, apps, sensores, internet de las cosas, etc. capaces de generar quintillones de bytes al día

BIG DATA

- El uso de Cloud Computing y Semantic Data para la gestión de Big Data en organizaciones es uno de los grandes desafíos en la evolución de la web

BIG DATA

- Big Data promete la generación de conocimiento, crear ventajas sostenibles y competitivas para las organizaciones

BIG DATA

- Se necesita Big Data cuando el análisis de información se ve afectado por el Volumen, la Variedad o la Velocidad en el procesamiento de datos:
  - Volumen: los datos son demasiado voluminosos para ser gestionados por nuestra infraestructura de datos actual

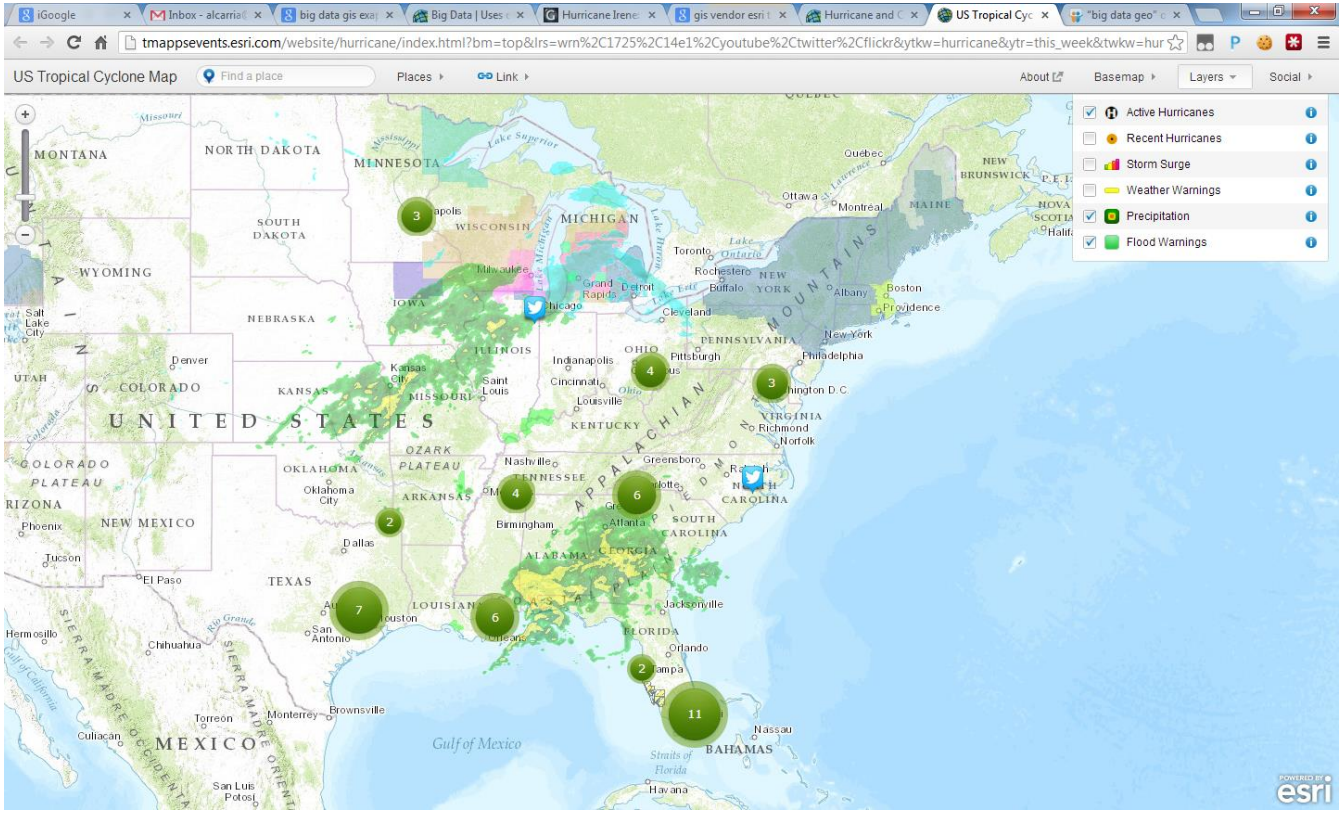
BIG DATA

- Variedad: hay demasiadas fuentes de datos de las que extraer información y en varios formatos (datos estructurados y no estructurados)

- Velocidad: necesitamos de manera ágil obtener conclusiones e información que nos ayude en tiempo real a tomar decisiones

BIG DATA

- Big Data con visualización geoespacial



BIG DATA

- Ejemplo: The New York Times utilizó, en 2007, 100 instancias de Amazon EC2 para procesar 4 TB de imágenes TIFF (guardadas en S3) para generar 11 millones de archivos PDF en 24 horas (240$)

- Amazon ha mejorado los servicios de procesamiento de Big Data con la introducción de EMR (Amazon Elastic MapReduce) en 2009

BIG DATA

- MapReduce
  - Introducido por Google en 2004 en el paper “MapReduce: Simplified Data Processing on Large Clusters”.
  - Objetivo: permitir la computación paralela sobre grandes colecciones de datos permitiendo abstraerse de los grandes problemas de la computación distribuida.
  - Usos: Geodesia, intersección de polígonos, carreteras, elevaciones

BIG DATA

- Esta técnica consiste en dos fases: Map y Reduce
- Las funciones Map y Reduce se aplican sobre pares de datos (clave, valor)

- Map toma como entrada un par (clave,valor) y devuelve una lista de pares (clave2,valor2). Se realiza en paralelo

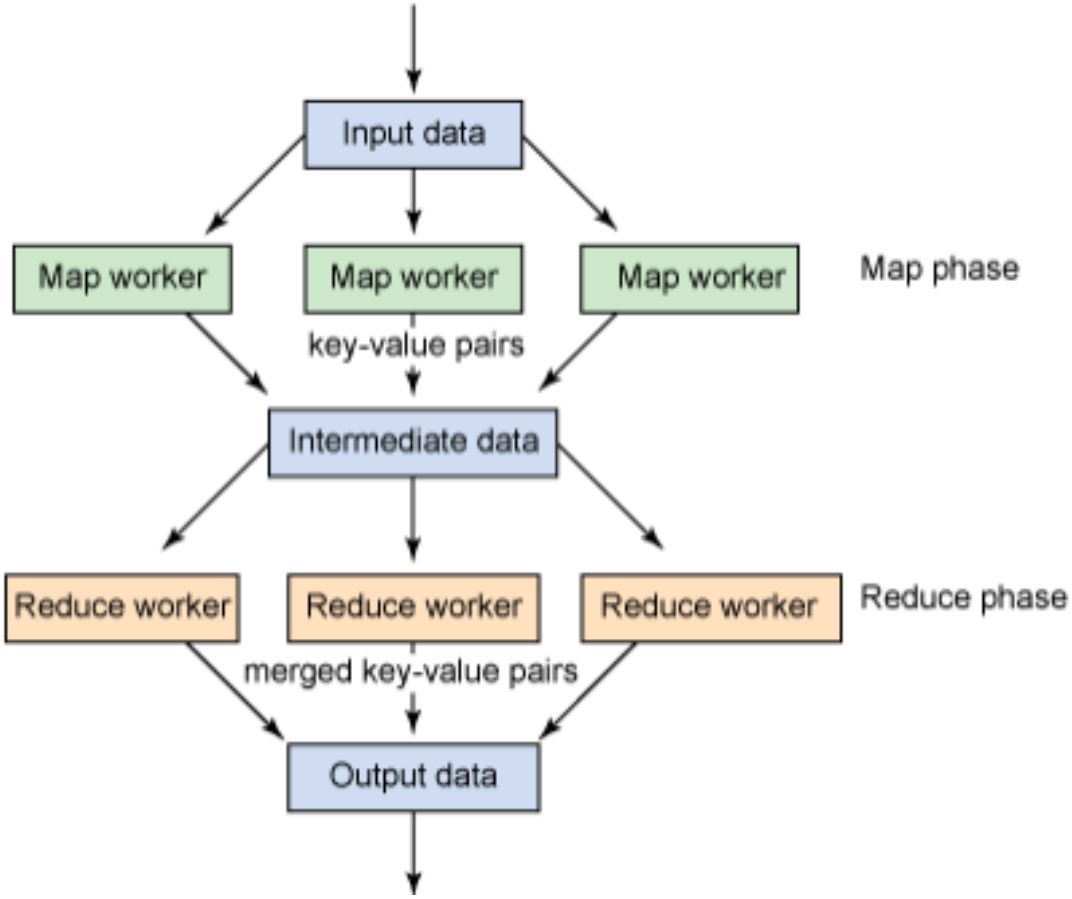
BIG DATA

- El framework MapReduce agrupa todos los pares generados con la misma clave

- Reduce se realiza en paralelo tomando como entrada cada lista de las obtenidas en el Map y produciendo una colección de valores.

BIG DATA

- Esquema de funcionamiento



BIG DATA

- Explicación del MapReduce: Contador de palabras

1 Nuestros datos

```
(3414, 'the cat sat on the mat')
(3437, 'the aardvark sat on the sofa')
```

2 Los pasamos al Map

```
map(String input_key, String input_value)
    foreach word w in input_value:
        emit(w, 1)
```

3 El Map genera

```
('the', 1), ('cat', 1), ('sat', 1), ('on', 1),
('the', 1), ('mat', 1), ('the', 1), ('aardvark', 1),
('sat', 1), ('on', 1), ('the', 1), ('sofa', 1)
```

4 El framework (Hadoop) agrupa los datos con la misma clave

```
('aardvark', [1])
('cat', [1])
('mat', [1])
('on', [1, 1])
('sat', [1, 1])
('sofa', [1])
('the', [1, 1, 1, 1])
```

5 Los pasamos al Reduce

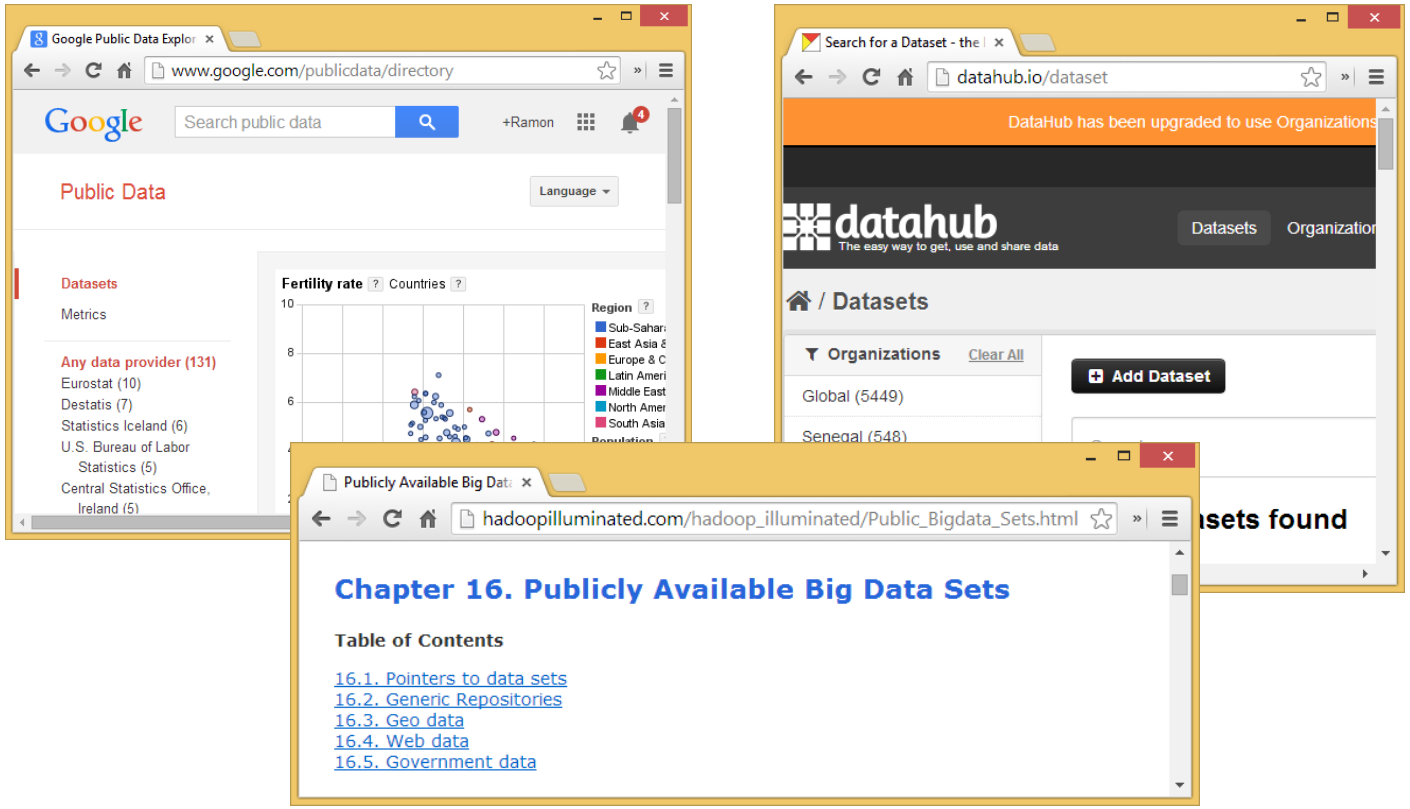
```
reduce(String output_key,
                Iterator<int> intermediate_vals)
    set count = 0
    foreach v in intermediate_vals:
        count += v
    emit(output_key, count)
```

6 Que genera el resultado

```
('aardvark', 1)
('cat', 1)
('mat', 1)
('on', 2)
('sat', 2)
('sofa', 1)
('the', 4)
```

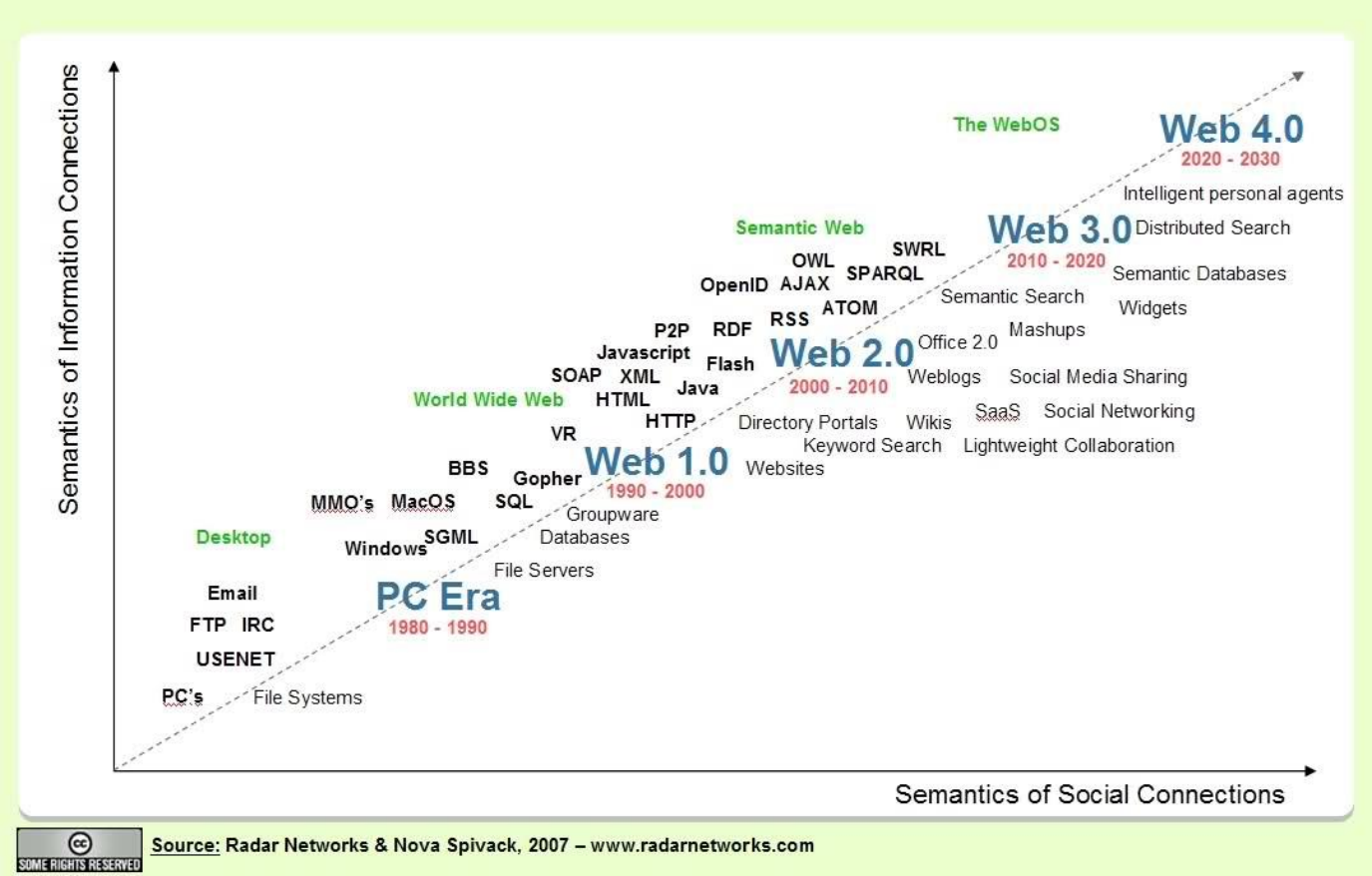
BIG DATA

- Amazon Web Services



SEMÁNTICA

- Evolución de la web



Source: Radar Networks & Nova Spivack, 2007 – www.radarnetworks.com

SEMÁNTICA

- Destinada a añadir significado a la web
- No existe total consenso de sus características:
  - Aumento de la interactividad y de la movilidad son dos factores que muchos señalan como decisivos en esta nueva etapa de la Web.
  - Web 3D, una web centrada en los medios de comunicación, una Web extendida, una gran base de datos presenta como páginas Web, o una combinación de todos ellos (Metz, 2007; Murugesan, 2007)

SEMÁNTICA

- Todavía no se ha producido este cambio

- Problema de la Web Actual:
  - El significado de la web no es comprensible por máquinas

SEMÁNTICA

- **Misión** → "convertir el contenido web actual en un contenido legible por una máquina"

¿Sería posible representar la información de forma que los ordenadores fueran capaces de interpretarla y ayudarnos de una manera automática a realizar nuestras búsquedas de una forma más precisa y relevante?

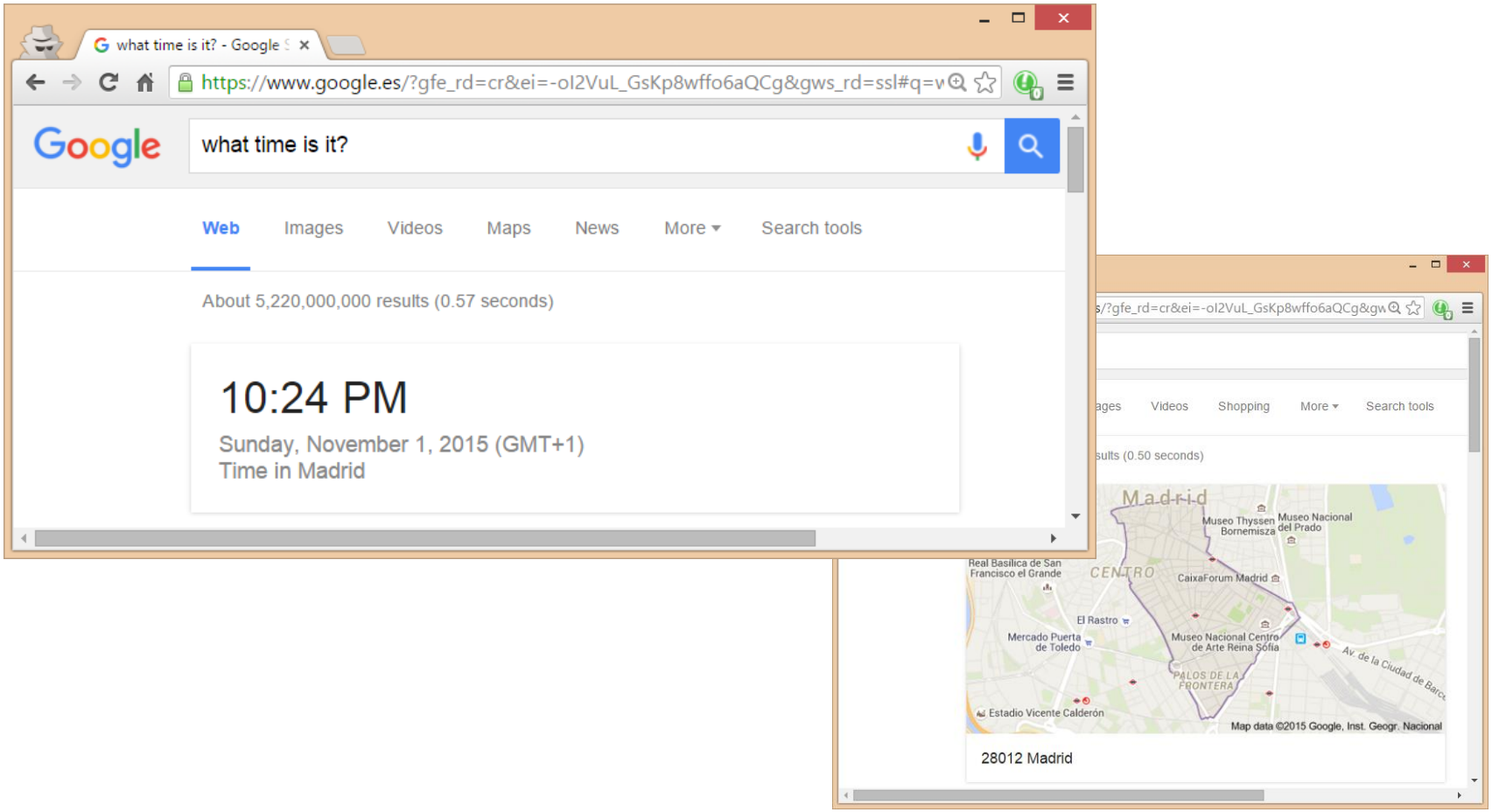
SEMÁNTICA

- Características:
  - Importancia al significado y a la relación de los datos.
  - Mayor capacidad de interacción entre los sistemas informáticos
  - Menor mediación de operadores humanos

SEMÁNTICA

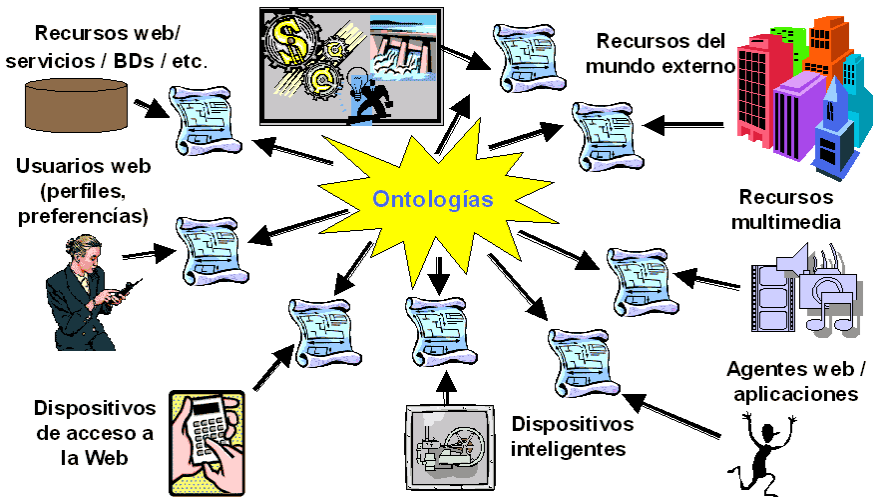
- Ejemplos:
  - Buscas una película y automáticamente obtienes enlaces al trailer, información para comprar una entrada de cine, información de los autores
  - Recuperar una canción presentando al motor de búsqueda un pequeño fragmento

SEMÁNTICA


what time is it? - Google
https://www.google.es/?gfe_rd=cr&ei=-oI2VuL_GsKp8wffo6aQCg&gws_rd=ssl#q=v
Google
what time is it?
Web
Images
Videos
Maps
News
More
Search tools
About 5,220,000,000 results (0.57 seconds)
10:24 PM
Sunday, November 1, 2015 (GMT+1)
Time in Madrid
Videos
Shopping
More
Search tools
Madrid
Museo Thyssen Bornemisza
Museo Nacional del Prado
Real Basílica de San Francisco el Grande
CENTRO
CaixaForum Madrid
El Rastro
Mercado Puerta de Toledo
Museo Nacional Centro de Arte Reina Sofía
PALOS DE LA FRONTERA
Estadio Vicente Calderón
Map data ©2015 Google, Inst. Geogr. Nacional
28012 Madrid

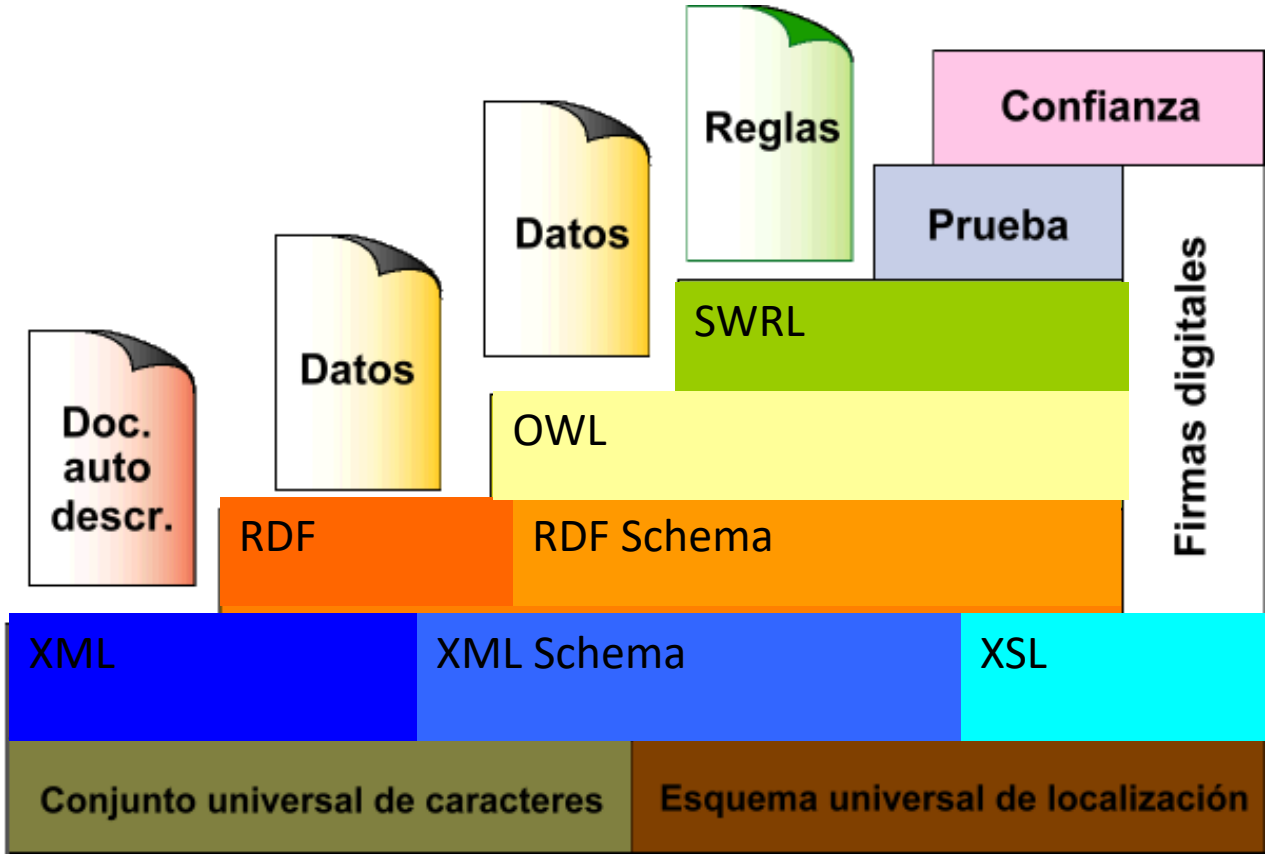
SEMÁNTICA

- En febrero de 2004, el World Wide Web Consortium (W3C) publicó las recomendaciones para el RDF y el Ontology Web Language (OWL). OWL describe la relación de cada uno de los componentes de la Web Semántica



SEMÁNTICA

- Caracterísiticas

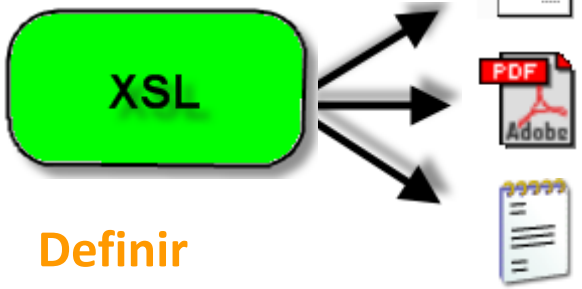


SEMÁNTICA

- Capa sintáctica
  - XML: Estándar de representación, metalenguaje para el intercambio de datos/información en la web en forma de documentos estructurados
  - XML Schema, restringe la estructura de documentos XML

Descripción normalizada

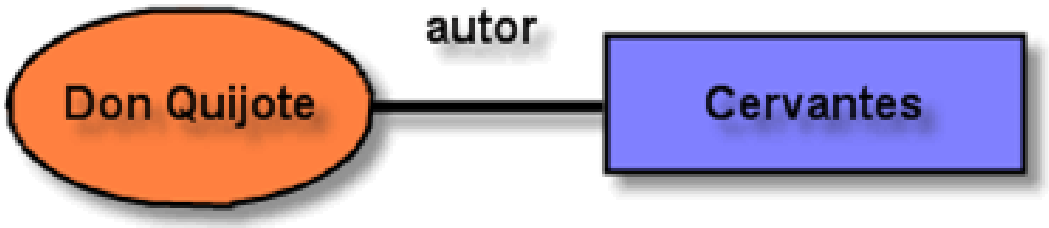
Definir y validar contenido



Definir formato

SEMÁNTICA

- Capa semántica
  - RDF (Resource Description Framework) es un modelo de datos que hace referencia a objetos y sus relaciones
  - RDF Schema, vocabulario para definir propiedades y clases de recursos RDF



SEMÁNTICA

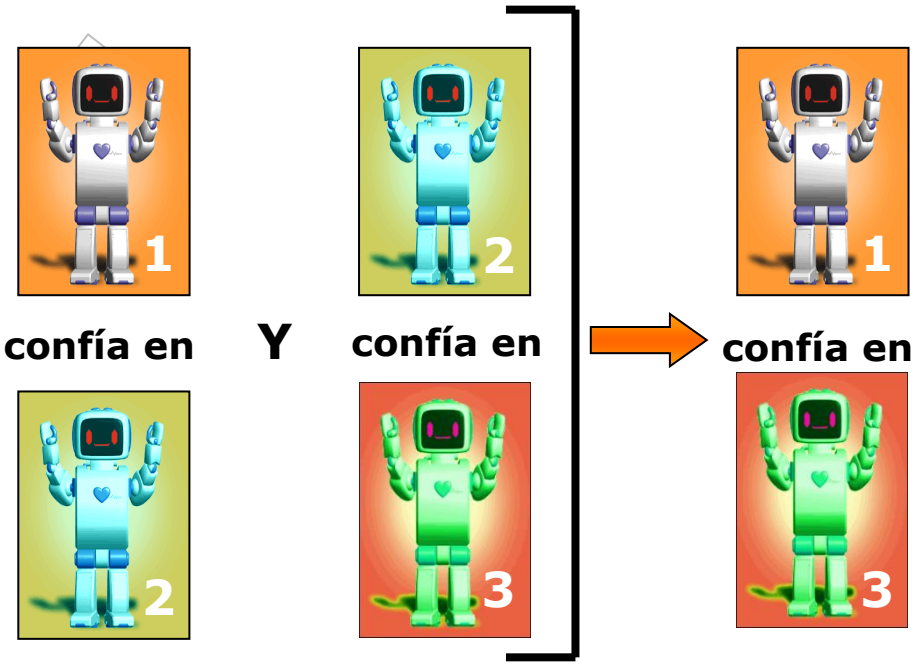
- Capa Ontológica
  - Ontología: Formulación de un exhaustivo y riguroso esquema conceptual dentro de uno o varios dominios dados; con la finalidad de facilitar la comunicación y el intercambio de información entre diferentes sistemas y entidades
  - OWL, añade más vocabulario que RDFS, relaciones entre clases, cardinalidad, igualdad …

SEMÁNTICA

- Capa lógica
  - Definición de reglas que permitan a los agentes operar con las inferencias obtenidas a partir de las ontologías
  - SWRL (Semantic Web Rule Language)
    - Definición de políticas de préstamo
    - Creación de sistemas de recomendación

SEMÁNTICA

- Capas de prueba y confianza
  - Prueba de que las respuestas a peticiones sobre la Web semántica son correctas
  - Confianza en las fuentes de datos



SEMÁNTICA

- Ejemplo de RDF

- **Formato RDF/XML:**

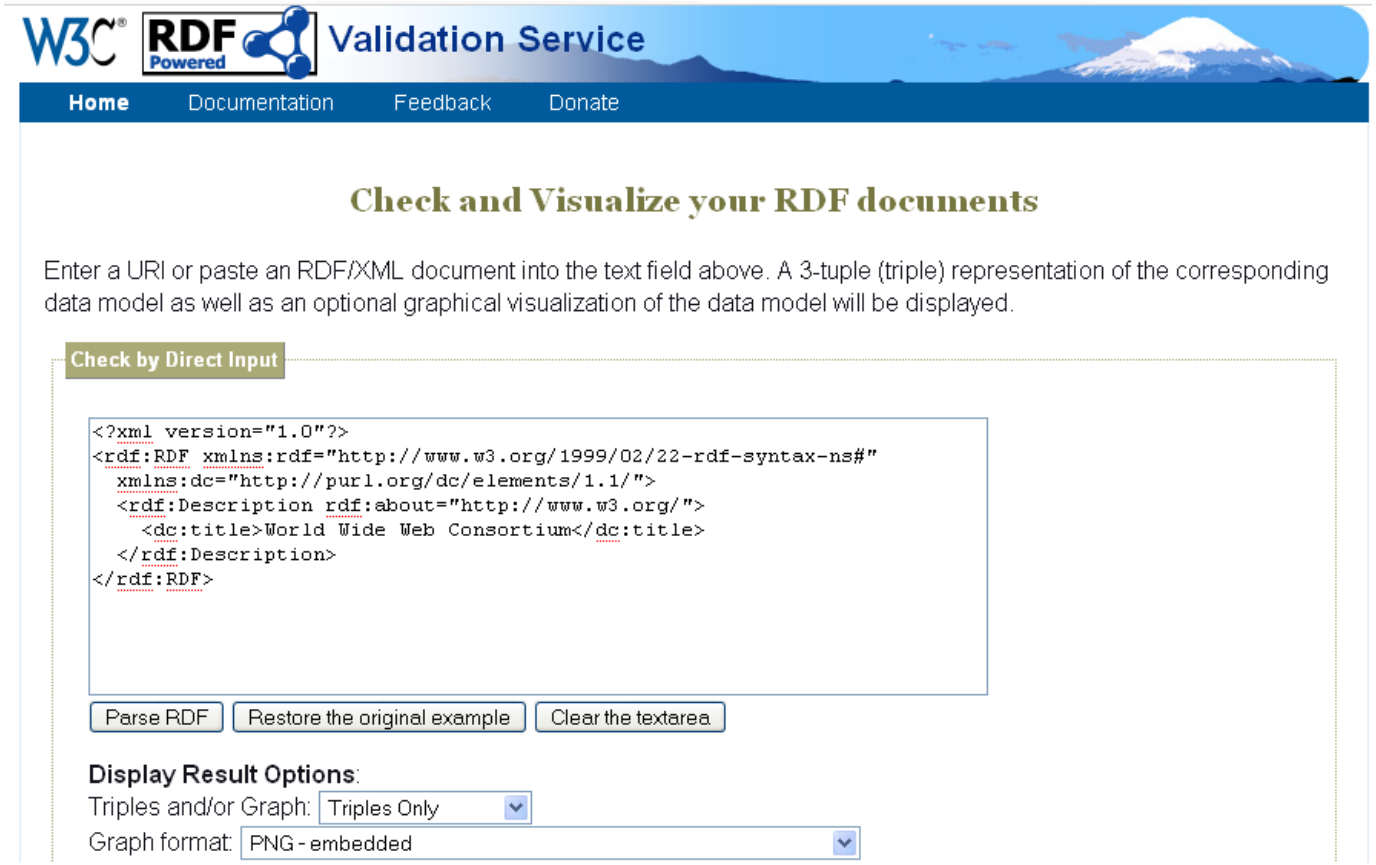
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos/"
    xmlns:edu="http://www.example.org/">
    <rdf:Description rdf:about="http://www.upm.es">
       <geo:lat>40.4519446</geo:lat>
       <geo:long>-3.7264568</geo:long>
       <edu:hasFaculty>
          <rdf:Bag>
             <rdf:li rdf:resource="http://www.etsit.upm.es" dc:title="Escuela de Teleco"/>
             <rdf:li rdf:resource="http://www.topografia.upm.es" dc:title="Escuela de Topografía"/>
             </rdf:Bag>
       </edu:hasFaculty>
    </rdf:Description>
 </rdf:RDF>
```

- **Formato: N3/Turtle:**

```
1: @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2: @prefix dc: <http://purl.org/dc/elements/1.1/> .
3: @prefix geo: <http://www. w3.org/2003/01/geo/wgs84_pos#> .
4: @prefix edu: <http://www.example.org/> .
5: <http://www.upm.es> geo:lat "40.4519446" ; geo:long "-3.7264568" .
6: <http://www.upm.es> edu:hasFaculty <http://www.etsit.upm.es> .
7: <http://www.etsit.upm.es> dc:title “Escuela de Teleco" .
8: <http://www.upm.es> edu:hasFaculty <http://www.topografia.upm.es> .
9: <http://www.etsit.upm.es> dc:title “Escuela de Topografía" .
```

SEMÁNTICA

- Validación RDF



SEMÁNTICA

- “Vocabulario” se ha convertido en una denominación genérica para hablar de estructuras o conjuntos de elementos normalizados en Internet

- En el ámbito de la SW: se ha formalizado la noción de "ontología“ como sinónimo de vocabulario

SEMÁNTICA

- Vocabularios:
  - Dublin Core (http://purl.org/dc/elements/1.1)
  - vCard-RDF (http://www.w3.org/TR/vcard-rdf)
  - Relationship ((http://vocab.org/relationship/)
  - Geo RDF (http://www.w3.org/2003/01/geo/)
  - Calendar RDF (http://www.w3.org/2002/12/cal/ical.rdf)
  - FOAF (http://xmlns.com/foaf/0.1/)

SEMÁNTICA

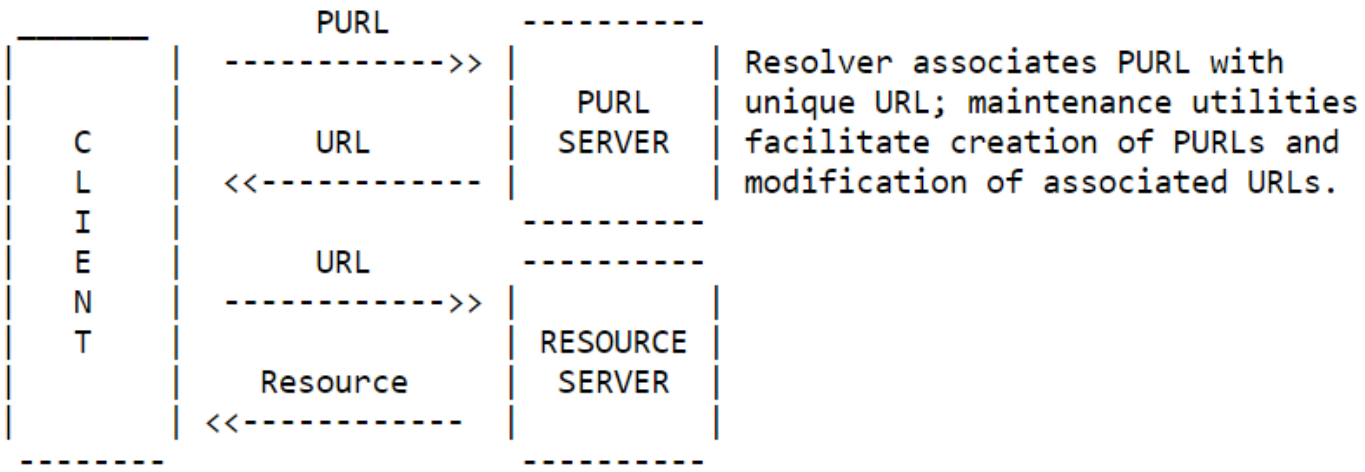
- Los vocabularios debe ser **persistentes**
- Para evitar que la URL del vocabulario cambie se utiliza el ***PURL (Persistent Uniform Resource Locator)***
- El localización PURL está asociado con la URL actual del vocabulario de forma que puede redirigir las peticiones.

SEMÁNTICA

- Los vocabularios debe ser **persistentes**
- Para evitar que la URL del vocabulario cambie se utiliza el ***PURL (Persistent Uniform Resource Locator)***
- El localización PURL está asociado con la URL actual del vocabulario de forma que puede redirigir las peticiones.

SEMÁNTICA

- https://archive.org/services/purl/



SEMÁNTICA

- XML: Provee una sintaxis elemental para las estructuras de contenidos dentro de documentos.
- XML Schema: Es un lenguaje para proporcionar y restringir la estructura y el contenido de los elementos contenidos dentro de documentos XML

SEMÁNTICA

- RDF: Es un lenguaje simple para expresar modelos de los datos, que refieren a los objetos “recursos” y a sus relaciones. Un modelo en RDF se puede representar en sintaxis de XML

SEMÁNTICA

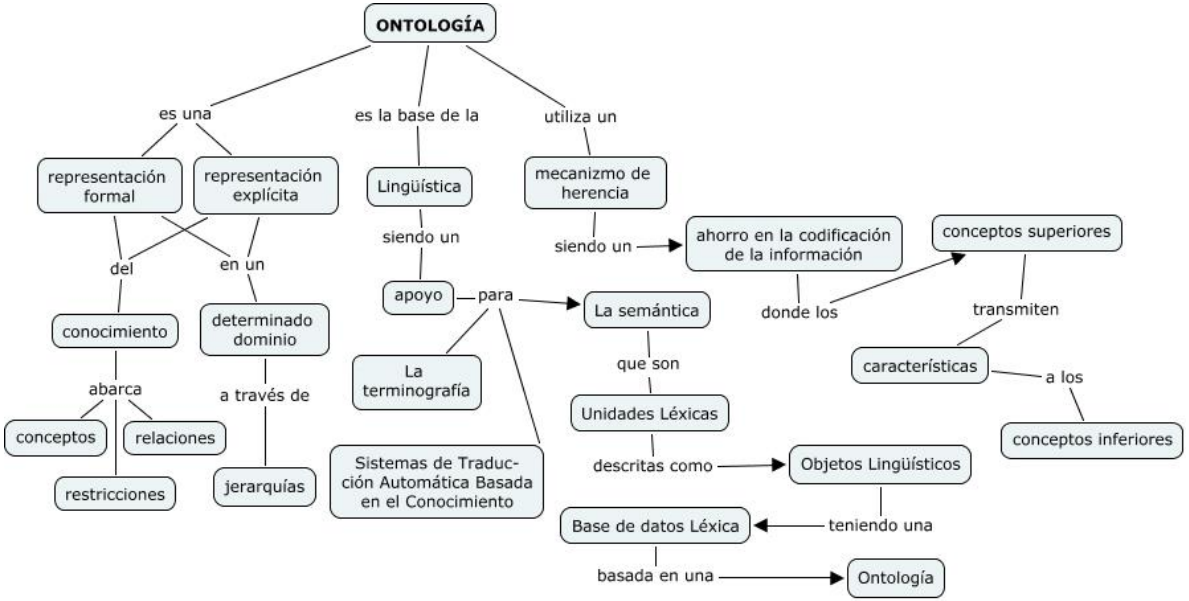
- RDF Schema: Es un vocabulario para describir propiedades y clases de recursos RDF-based, con semántica para generalizar-jerarquías de las propiedades y clases.
- OWL: Es un mecanismo para desarrollar temas o vocabularios específicos en los que podamos asociar esos recursos

SEMÁNTICA

- Una ontología define conceptos de un dominio y relaciones entre ellos
  - Los bloques básicos que componen el diseño de una ontología son:
    - clases o conceptos
    - propiedades de cada concepto describiendo varias características y atributos del concepto
    - restricciones sobre las propiedades

SEMÁNTICA

- Una ontología junto con las instancias de sus clases individuales constituyen un knowledge base



Fuente:
http://www.hipertext.net/web/pag220.htm#3.1
http://es.wikipedia.org/wiki/Ontolog%C3%ADa_(inform%C3%A1tica)

SEMÁNTICA

- Una ontología difiere de un esquema XML en que es una representación de conocimiento, no un formato de mensaje
- La principal ventaja de una ontología escrita en OWL es que hay disponibles herramientas que pueden razonar sobre ella
- La sintaxis de intercambio de información en OWL es normalmente RDF/XML

SEMÁNTICA

- Las ontologías Web son distribuidas
- Pueden ser importadas y extendidas para crear ontologías derivadas
- Se pueden alinear unas ontologías con otras

SEMÁNTICA

Supongamos el siguiente modelo RDF:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<http://www.ipina.org/> foaf:author
<http://www.ipina.org/osgi/> .
<http://www.deusto.es/dipina/> foaf:author
<http://www.deusto.es/dipina/ajax/> .
<http://www.eside.deusto.es/dipina/> foaf:author
<http://paginaspesonales.deusto.es/dipina/> .
```

Aunque pertenecen al mismo autor, no están relacionadas entre ellas, con la ayuda de OWL podemos mapear estas URIs

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
<http://www.deusto.es/dipina/> owl:sameAs
<http://www.ipina.org/> .
<http://www.eside.deusto.es/dipina/> owl:sameAs
<http://www.ipina.org/> .
```

Si mezclamos ambos modelos y ejecutamos un razonador podríamos responder a “dime todo lo que ha escrito “<http://www.ipina.org>”:

```
<http://www.ipina.org/osgi/>,<http://www.deusto.es/dipina
/ajax/> y <http://paginaspesonales.deusto.es/dipina/>
```

SEMÁNTICA

- Los microformatos intentan ser útiles principalmente a las personas y en segundo lugar a las máquinas; para ello aprovechan características de [X]HTML para añadir información semántica en una sección de código [X]HTML.

- Análisis de contenido “de autor” en la Web 2.0.

SEMÁNTICA

- Los microformatos ofrecen soluciones sencillas a problemas de representación de información concretos en la Web (p. ej.: cómo codificar la información personal de una tarjeta de visita, cómo codificar un evento, un vuelo, etc.