USID07
Simplicity is the ultimate sophistication
Leonardo da Vinci

De la web sintàctica a la web semàntica

Lluís Codina (Universitat Pompeu Fabra)

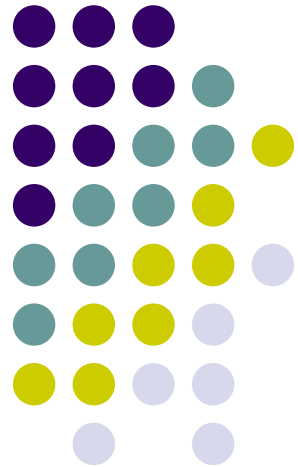
Resum

Si s'acompleixen els pronòstics del World Wide Web Consortium, l'organisme que promou els estàndards més importants d'Internet, el futur web no serà com l'actual, el contingut del qual solament poden interpretar les persones, sinó que estarà format per llocs que els ordinadors també podran "interpretar" i fer "raonaments" sobre el contingut semàntic de les pàgines. Aquest és un escenari realista per al qual hem de preparar-nos de cara als propers anys o hem de relegar aquesta visió a un futur llunyà? En aquesta conferència es presentaran les bases d'aquest nova web que, en part, ja tenim amb nosaltres; s'examinaran els canvis principals que aportarà, i s'examinaran les conseqüències a curt i mitjà termini per als professionals de la documentació implicats amb la gestió i/o publicació d'informació digital en línia..

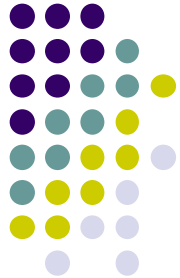
De la Web Sintáctica a la Web Semántica

Lluís Codina (UPF)
www.lluiscodina.com

4ª Jornada de Usabilidad en Sistemas de Información Digital
Barcelona, Mayo 2007

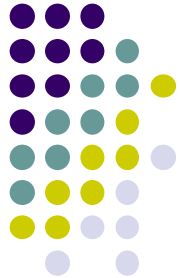


Qué es la Web Semántica



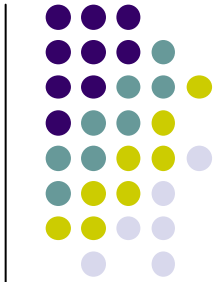
- Un conjunto de normas (recomendaciones) del W3C
- *La visión*: una Web cuyo contenido puedan interpretar los ordenadores (¿Inteligencia Artificial?)
- *La motivación*: Una infraestructura para el comercio electrónico y los servicios web
- *Un subproducto*: ¿Una infraestructura para la gestión del conocimiento?

Definiciones de la Web Semántica (1)



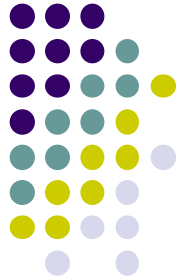
- W3C:
  - The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries (…). It is based on the Resource Description Framework (RDF)
- Wikipedia:
  - La Web semántica es la idea de añadir metadatos semánticos a la World Wide Web

Definiciones de la Web Semántica (2)



- W3C:
  - La Web Semántica es una Web extendida, *dotada de mayor significado* en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a *una información mejor definida*. Al dotar a la Web de *más significado* y, por lo tanto*, de más semántica,* se pueden obtener soluciones a problemas habituales en la *búsqueda de información.*

El síndrome del elefante o las (al menos) tres almas de la WS

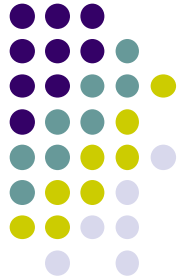


- La visión de la Inteligencia Artificial (IA) > *Ontologías*
- La visión de la bases de datos (SGBD) o “del procesamiento robusto” > *XML + Metadatos*
- La visión de los servicios*:* la web semántica no es “solo” para encontrar información

Una contradicción esencial

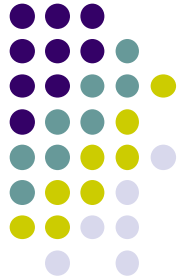
- El punto común: una web que permita *razonar* a los ordenadores, realizar *inferencias* y tomar *decisiones*.
- El problema esencial:
  - Los ordenadores son máquinas sintácticas y la mera sintaxis no produce semántica. Dicho de otro modo:
    - La hipótesis del sistema de símbolos físicos (A. Newell y Herbert A. Simon) vs.
    - La hipótesis de la habitación china (John Searle)

Componentes principales de la WS



- *XML*: marcado semántico
- *RDF*: sistema común para expresar metadatos
- *OWL*: sistema común para expresar ontologías
- *Agentes de usuario*: para unir todo lo anterior al servicio del internauta

Infraestructura necesaria



- Servidores y sitios web:
  - Con marcados semántico vía XML (p.e. XHTML)
  - Con metadatos
  - Eventualmente: con ontologías asociadas
- Agentes de usuario:
  - Navegadores no “tolerantes”
  - Capaces de interpretar metadatos
  - Capaces de interpretar ontologías o de invocar las aplicaciones necesarias

XML

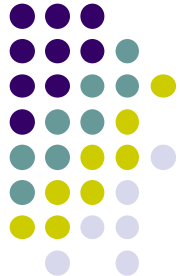
- XML: lenguaje para definir lenguajes con etiquetas semánticas (y no de presentación). Ejemplo:

  <autor>Umberto Eco**</autor>**

  vs. ****Umberto Eco****

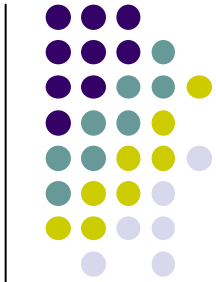
- XML Schema:
  - Especificación para asignar tipos de datos, dominios, rangos de valores y restricciones a las etiquetas XML

RDF



- ***Resource Description Framework:***
  - W3C: The *Resource Description Framework* (RDF) integrates a variety of applications from *library catalogs and world-wide directories to syndication and aggregation of news*, software, and content to personal collections of music, photos, and events using XML as an interchange syntax. The RDF specifications provide a *lightweight ontology system* to support the exchange of knowledge on the Web

RDF - Metadatos

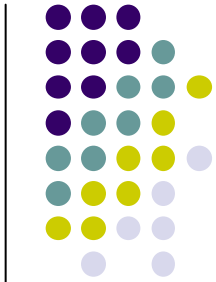


- RDF: Un sistema de descripción de entidades (recursos) con una base lógico/lingüística
- RDF relaciona recursos con propiedades y valores
- Proporciona un sistema común de expresión de metadatos

Ejemplo RDF

En modo nativo (gráfico):

En modo serializado (RDF/XML)

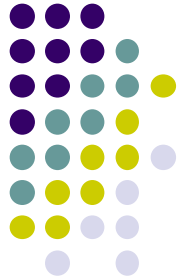


...

```
<rdf:Description rdf:about="http://w3.org/">
   <dc:title>World Wide Web Consortium</dc:title>
<rdf:Description>
```

...

RDF vs SGBD

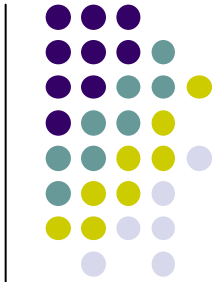


- Modelo RDF:
  - Un recurso (*sujeto*) tiene una propiedad (*predicado*) con un determinado valor (*objeto*)
  - Ejemplo: El *libro ID123* tiene un *título* y el valor del título es *Romeo y Julieta*
- Equivale a:
  - Una *entidad* (registro) tiene un *atributo* (campo) con un determinado *contenido* (valor)
- O bien:
  - Recurso=*Fila*; Propiedad=*Columna*; Valor=*Valor*

OWL

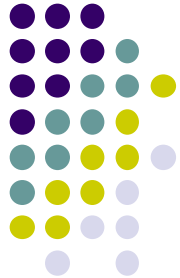
- ***OWL: Web Ontology Language***
  - OWL builds on RDF and RDF Schema and *adds more vocabulary for describing properties and classes*: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

OWL - Objetivos



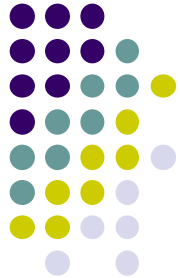
- OWL (…) add the following capabilities to ontologies:
  - Ability to be distributed across many systems
  - Scalability to Web needs
  - Compatibility with Web standards for accessibility and internationalization
  - Openess and extensiblility

Otra forma de verlo: la WS como una base de datos



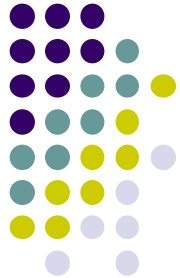
- *Cada unidad* significativa de texto (palabra, frase, oración, párrafo, página) está delimitada (marcada) mediante *etiquetas* XML (cada unidad es un elemento*)*
- *Cada elemento* tiene asociado un *tipo de dato (*vía *schemas)*
- *Cada documento,* como un todo, contiene (o está asociado a) un conjunto de *metadatos.* Incluso puede tener metadatos a *nivel de elemento (*vía *RDFa)*
- *Resultado:* la Web como una gran base de datos descentralizada, distribuida y no coordinada (registros formados por campos + diccionario de datos + descriptores)

La WS realmente existente: aplicaciones XML



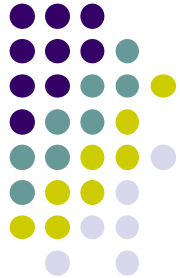
- Suites ofimáticas
- SGBD
- Editores de sitios web
- Navegadores

Ejemplos de software/iniciativas para la WS



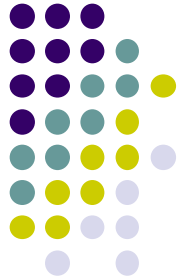
- **XML**
  - Altova XML Spy
  - Altova Semantic Works
  - Amaya/Anotea
- **Metadatos**
  - Dublin Core
- **RDF**
  - Protégé
  - Smore

¿Dónde está la WS?



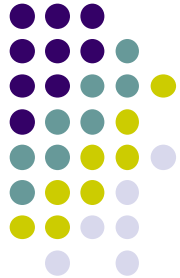
- **No está:**
  - En los motores de búsqueda actuales (evitan expresamente los metadatos)
  - En bases de datos (p.e. no está en Scirus, ni en ISI, etc.)
  - En la mayor parte de la web “real”
- **Empieza a estar:**
  - En los sitios web que usan estándares de manera estricta y aplican marcados semántico
  - En algunos repositorios (*e-prints*, *pre-prints*, etc.)
  - En un reducido (pero selecto) número de sitios web relacionados con la Administración y/o con iniciativas de carácter científico o cultural

Conclusiones (1): ¿Qué está aportando la WS?



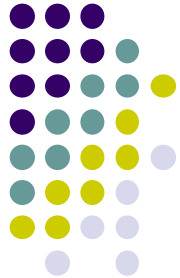
- **Actualmente:**
  - Un nuevo formato universal de datos: XML
  - Una fuerte impulso al uso de estándares Web y un redescubrimiento del marcado semántico (HTML y HTML 5)
  - Un renovado debate sobre el uso, definición y alcance de los metadatos
  - Un nuevo formato universal para expresar metadatos: RDF con aplicación a tesauros y lenguajes documentales
- **En el futuro:**
  - ¿Servidores de ontologías?
  - ¿Nuevos sistemas de búsqueda y acceso a la información?
  - ¿Una nueva generación de repositorios, bibliotecas digitales y sistemas de información?

Conclusiones (2): Paradojas



- **Una web más alejada del público:**
  - Hay que tener conocimientos más técnicos para desarrollar páginas web:
    - Declaraciones DOCTYPE, CSS, elementos depreciados, separación de contenido, y formato, etc.
  - Resultado: es necesario conocer más detalles en lugar de que queden ocultos, contradiciendo la evolución “natural” de la informática
- **Nuevos estándares *no estándares*:**
  - HTML 5 > un HTML que no es del W3C
  - Microformatos > metadatos que no son del W3C
- **Mayores dificultades para desarrollar sitios Web:**
  - Se deben añadir metadatos a nivel de sitio, página y elemento
  - Marcado semántico: mayor número de elementos, etiquetas y atributos
- **Falta de actores con alicientes claros:**
  - ¿A quién beneficia aplicar los estándares de la WS?
  - Algunos problemas que afronta la WS están solucionados por otras vías (p.e. el análisis de enlaces)

¿Qué podemos hacer?



- El objetivo de la Web Semántica es magnífico. Es la reedición para el Siglo XXI del proyecto del *Acceso Universal al Conocimiento*. Propuestas:
  - Dar soporte al uso de estándares del W3C (XML, XHTML) > Nuevas páginas o nuevos sitios + Conversión retrospectiva de los ya existentes
  - Utilizar el marcado semántico *ya disponible* en (X)HTML, tanto en forma de elementos (address) como de atributos (title)
  - Usar aplicaciones y modelos de datos que utilicen XML
  - Expresar metadatos mediante RDF (RDFa, RDF/DC, etc.)
  - Concebir las ontologías como nueva frontera de la semántica documental, estudiar sus posibilidades y, eventualmente, promocionar su aplicación

Referencias

- D. Fensel *et. al*. *Spinning the semantic web.* Cambridge: MIT, 2005
- G. Antonou; F.v. Harmelen. *A semantic web primer*. Cambridge: MIT, 2004
- L. W. Lacy. *OWL: Representing information using the Web Ontology Language*. Ann Arbor: Trafford, 2004
- D. R. Miller; K. S. Clarke. *Putting XML to work in the library*. Chicago: ALA, 2004
- J. Tramullas (coord.) *Tendencias en documentación digital*. Gijón: TREA, 2006