LA WEB SEMANTICA

claudio gutierrez
Centro de Investigaciones de la Web, CIW
Departamento de Ciencias de la Computación
Universidad de Chile
cgutierr@dcc.uchile.cl

Agenda

- La Web: puntos de vista
- Dos ejemplos
- Web Futura: Estándares

Web: dos puntos de vista

- Consumidor: acceso rápido y barato a múltiples y variadas fuentes de información.
- Productor: posibilidad de "exponer" sus puntos de vista sin restricciones, a millones de personas, en forma continua, en casi cualquier

  "formato" (texto, imágenes, audio, video, etc.)

Web: más puntos de vista

- gigantesca base de datos distribuída
- gran red de comunicaciones
- biblioteca (casi) infinita
- gran foro virtual
- primer ejemplo real de cerebro artificial

Web actual: datos

- Tamaño: orden de los Petabytes (miles de Terabytes) (Biblioteca Congreso USA: aprox. 20 TB en libros)
- Cientos de miles de millones de documentos individuales (orden de 10^{12})
- Más del 95 % información gratis
- Crecimiento: servidores doblan cada dos años

Resumen: MUUUY GRANDE...

Web actual: más datos ...

- Heterogeneidad de tipos de información
  (texto, video, sonido, música, gráfica, etc.)
- Heterogeneidad de formatos
  (Para cada tipo anterior, diversos formatos)
- Heterogeneidad idiomática
- Heterogeneidad semántica
  (objetos tienen diferentes significados para diferentes usuarios)

Web actual: problemas prácticos

- Dificultad para encontrar la información que se desea
  Ej: Lista de las orquestas con que cantó Julio Sosa
- Respuestas poco precisas
  Ej: Varios cientos de miles de páginas sobre Borges
- Falta de automatización
  Ej: dirección postal de las Universidades latinoamericanas
- Heterogeneidad de interfaces
  (¡Sin comentarios!)

... y los problemas crecen con la web exponencialmente

Web actual: problemas técnicos

- No hay catalogación global
- Débil discriminación sobre la información que existe
- Deficiente integración de información
- Automatización de las funciones anteriores inexistente
- Centros de confianza

Web actual: posibles soluciones

1. Apostar a éxitos de IA (y rezar mucho...)
2. Considerar la Web como una gran Base de Datos
   y adaptar esas tecnologías
3. Aprender de nuestros viejos amigos los bibliotecarios
   ... pero automatizar sus funciones
4. Descentralizar

1. Exitos de IA

Aprovechar:

- Representación del conocimiento
- Deducción automática
- Procesamiento lenguaje natural

No esperar que inteligencia artificial resuelva lo que no puede la inteligencia natural.

2. Web actual como Base de Datos

La Web es una gigantesca base de datos distribuída

Detalles...

- El administrador cargó los datos antes de tener los esquemas
- No fijó los permisos para crear y modificar
- No especificó formato ni tipos de datos a usar

3. Las lecciones de la biblioteca

- Metadatos
- Metadatos
- Metadatos

4. Descentralizar

- Esperanto no funciona (sí los diccionarios...)
- La información actualizada está en los nodos
- La Web es una red, no una pirámide
- Compartir y facilitar es más eficiente que monopolizar e imponer (aunque probablemente menos rentable...)

La Web del Futuro

AUTOMATIZAR el proceso de:

- Comprender la información
- Organizar la información
- Validar información
- Deducir nueva información

¿Que necesitamos para automatizar estos procesos?

¿Y quien nos ayudará?

La...W3C (Consorcio de la Web)

"Mission: W3C leads the evolution of the web, empowering individuals, increasing social and economic efficiency, and exploiting the power of computing in our everyday lives."

Ejemplo 1

Año 1936: el doctor Julio Medina posee una biblioteca personal con cientos de libros, que ocupan dos piezas grandes. El orden no es el fuerte del doctor. Así, cada vez que necesita cierta información, debe revisar cada libro, hasta encontrar lo que busca.

Problemas:

- no siempre encuentra lo que busca
- búsqueda toma mucho tiempo

Ejemplo 1 (continuación)

Solución: Catalogar los libros

- ¿Cómo referenciar cada libro de forma única?
  - Cuatro letras del autor + año (como nombrarlo)
  - Nro. de estante + correlativo (donde encontrarlo)
- ¿Qué alfabeto utilizar?
  - Latino
- ¿Cómo anotar la información de los libros?
  - En fichas almacenadas en tres cajas de madera según criterio. Cada ficha tiene la referencia a un libro.
- ¿Qué criterios utilizar para catalogar los libros?
  - autor, título y tema

Ejemplo 1 (continuación)

Año 1937: se muda a la casa contigua a la del doctor, un escritor ruso, Ivo Ronikoff, quien también tiene una gran biblioteca. Ivo es ordenado y su biblioteca está catalogada así:

- Referencias:
  - tres letras del título + año para nombrarlo
  - correlativo como ubicación
- Alfabeto: cirílico
- Información sobre los libros: anotada en gordos cuadernos tipo guía de teléfono.
- Criterios de catalogación: palabra clave, editorial y año.

Ejemplo 1 (continuación...)

Año 1938: Llega otro vecino, también un gran aficionado a los libros: el profesor Tsun Ming Fu, recién venido de Shangai.

...ni nos imaginemos como ordena su biblioteca...

Ejemplo 1: ¿Cómo compartir información?

El Dr. Medina, el Sr. Ronikoff y el profesor Ming Fu gustan compartir sus libros, pero nunca les fué fácil buscar información en las tres bibliotecas.

Problemas:

- Alfabetos diferentes
- Referencias diferentes
- Anotaciones diferentes
- Criterios de catalogación diferentes.

Ejemplo 1: ¿final feliz?

Año 2002: Ana Medina Ronikoff reprende a sus abuelos por tercos y los insta a "subirse a las nuevas tecnologías". Les propone usar:

- alfabeto: UNICODE
- referencias: URI
- Anotaciones: XML y RDF
- Criterios de catalogación: Ontologías

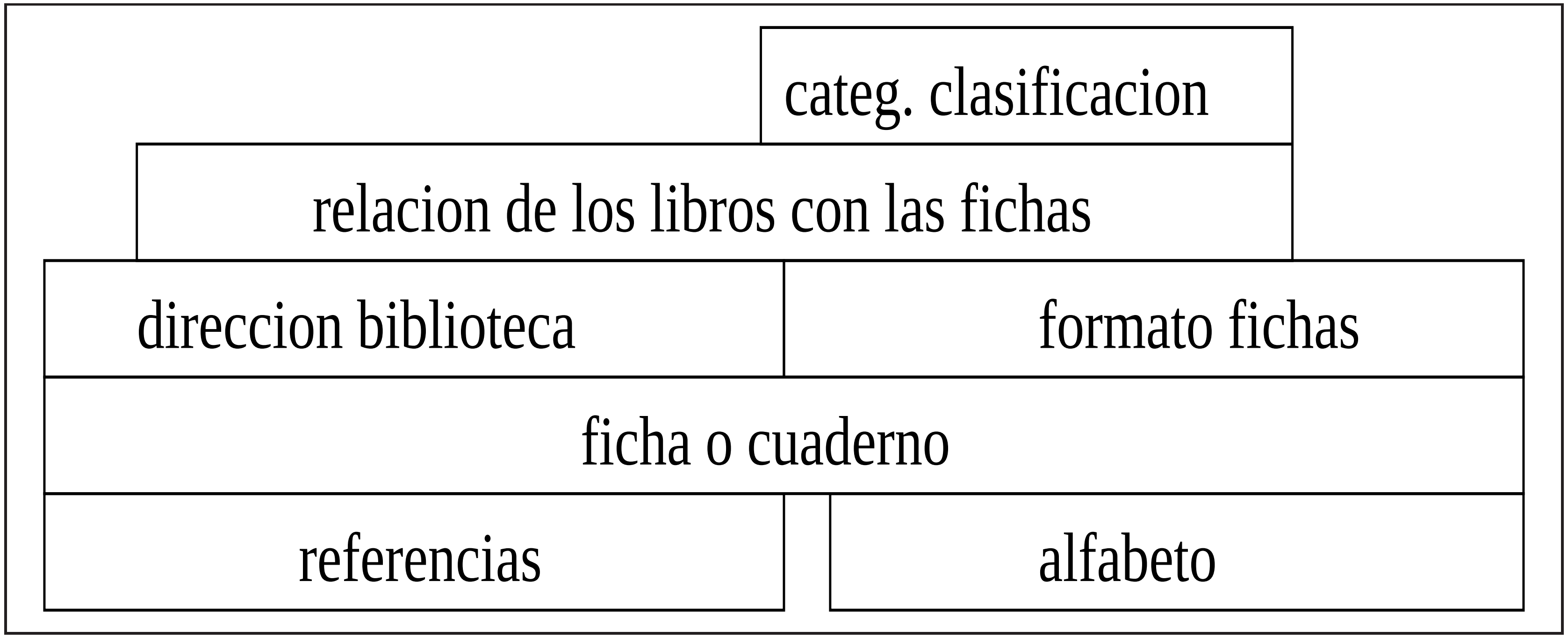
Ejemplo 1: Las dudas de los abuelos

Los abuelos se sienten insultados y replican:

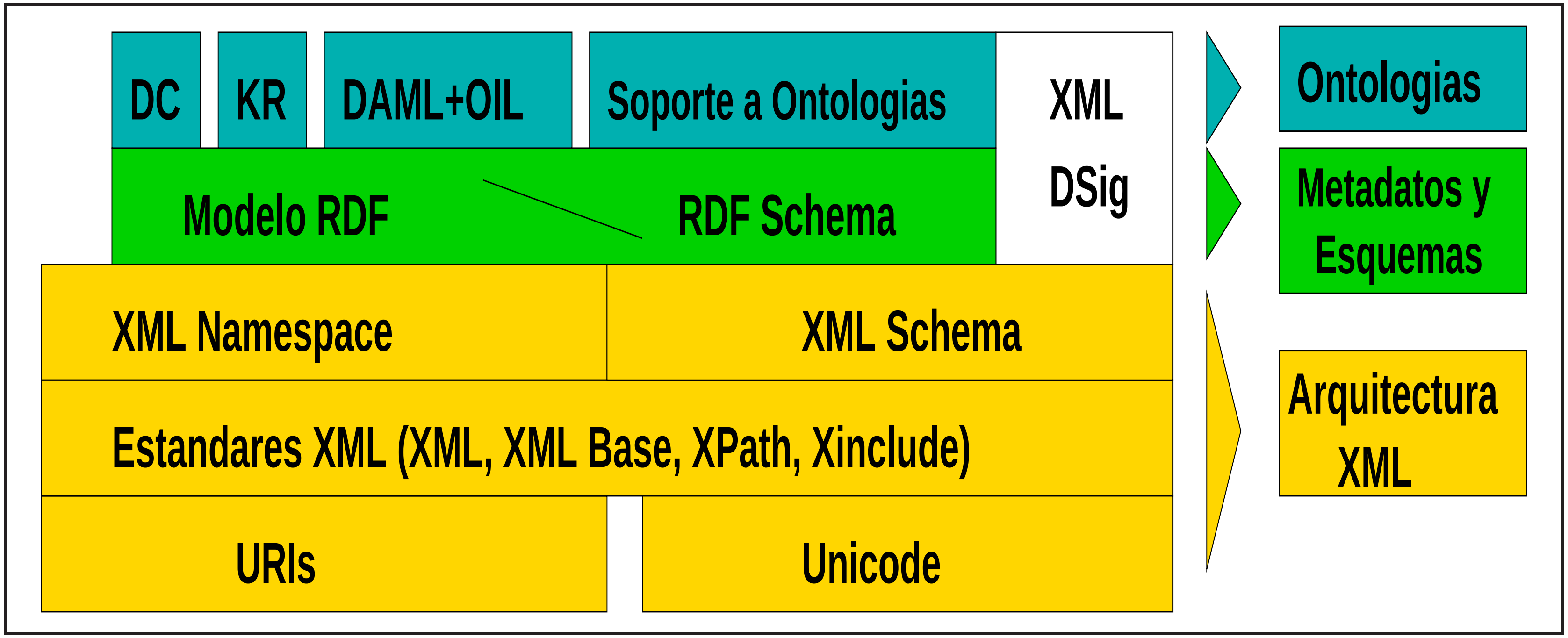
- El Dr. Medina: ¡tanta parafernalia!
- El escritor Ivo, algo más tecnologizado:
  ¿Por que mejor no usar simplemente HTML?
- El profesor Ming Fu, asiduo lector de Infoweek:
  ¿Y por que XML no nos basta?

Ejemplo 1 (continuación)

La Arquitectura de la solución de Ana Ronikoff



Arquitectura de la Web Semántica



Ejemplo 2: la Web actual

Año 2002: Para un importante ejecutivo de una gran empresa es muy costoso buscar información en Internet por el tiempo que le toma. Por ello delega esta responsabilidad a un asistente.

El asistente busca la información así: utiliza un browser y en sitios como Google o Altavista encuentra gran cantidad de enlaces. Los revisa y si tiene suerte, extrae la información solicitada por su jefe. El ejecutivo recibe la información por parte de su asistente algunas horas después.

- Aquí el asistente es un agente humano.

Ejemplo 3: ¿la Web futura?

Año 2010(?): Un niño pregunta a su asistente personal que lleva puesto (con un monóculo como pantalla y algo como un teclado en la manga de su chaqueta): "¿Quién era el entrenador cuando la roja le puso dos goles al campeón del mundo en el 2001?". El asistente personal levanta un proceso en background y busca sitios Web donde encontrar lo solicitado por el niño, quien en poco más de un minuto recibe la respuesta: Pedro García.

- Aquí el asistente es un agente digital.

(Lo que no sabía el agente es que ese mismo equipo saldría último en el mundial del 2002 ... nadie es perfecto!)

Agentes - Hoy y Mañana

Los asistentes de los ejemplos 2 y 3 son agentes que tienen las siguientes características:

- Entienden lo que se les pide buscar
- Comprenden el contenido de los sitios Web que visitan
- Validan si lo encontrado corresponde a lo que se le pidió buscar
- Deducen nueva información de la ya obtenida

Web Futura: estándares

Temas críticos:

1. Modo único de referenciar (URI)
2. Lenguaje estándar de intercambio (XML)
3. Creación, administración, mantención de metadatos
4. Lenguage estándar de metadatos (RDF)
5. Estandarización de vocabularios (Ontologías)
6. Y otras cositas más...
   (centros de confianza, capacidad de deducción, etc.)

URI

- URI = Uniform Resource Identifier
- Representa la identificación de un recurso por un nombre
- Pueden ser creadas por cualquier persona y usadas como se estime más conveniente
- No están limitadas a objetos que tienen dirección en la red u otros mecanismos computacionales de acceso (ej: mi perro, conceptos abstractos, etc.)
- Stock infinito de nombres: "Una URI es una cadena compacta de caracteres para identificar un recurso físico o abstracto".

Ventajas: Uniformidad de referencia, Universalidad de referencia, Descentralización (ej. freenet)

XML

Familia de tecnologías

- XSL: Presentación y transformación de documentos. (XSLT, XPath, XSL-FO).
- XMLSchema: definición de documentos, restricciones de integridad, etc.
- XQuery: lenguaje de consulta para documentos XML
- Xlink: estandariza hiperlinks

Ventajas de XML

- Universalidad via espacios de nombres
- Modularidad via espacios de nombres
- Versatilidad via datos semiestructurados
- Sólidos fundamentos via SGML
- Amplio soporte y distribución

Metadatos

- Datos estructurados acerca de datos
- Información descriptiva acerca de un objeto o recurso
- Ejemplo clásico: tarjetas de catálogos de bibliotecas (DCC, LCC, DC)
- Otros ejemplos:
  - Gobierno: GILS, Government Information Locator Service, ANZLIC (Australia New Zealand Land Information Council) AGLS, Australian Government Locator Service)
  - Datos geográficos y espaciales: CSDGM, FGDC tiene muchos más

Problemas con metadatos hoy día: No son extensibles globalmente; Modelos distintos e incompatibles

RDF

- Lenguaje de metadatos serializado en XML
- Provee interoperabilidad entre aplicaciones (agentes)
- Modelo robusto desde el punto de vista lógico
- Elaborado por W3C y participación de las comunidades de metadatos

Defecto: algo complicado ... no tanto

¿Que otra cosa nos falta?

- Firmas Digitales
- Lógica, Deducción, Confianza
- Lenguajes de Consulta
- Aplicaciones matadoras
- Varias otras cosas...