# Acuerdo de Bibliotecas Universitarias de Córdoba

Seminario

27 y 28 de septiembre de 2012

# Web semántica, Web 3.0 y entornos Cloud Computing, nuevos horizontes para bibliotecarios, documentalistas y archivistas

Mela Bosch

melabosch@hotmail.com

# Acuerdo de Bibliotecas Universitarias de Córdoba

Seminario

27 y 28 de septiembre de 2012

## Quinto encuentro: Motores de búsqueda y anotación semántica

**Temáticas: Recapitulamos modelos de datos. Mercado de motores de búsqueda. Tipos de buscadores. Anotación. Tipos de anotación. Tratamiento del lenguaje natural**

**Cierre: intercambio y puesta en común**
**Mi experiencia con motores de búsqueda**

Mela Bosch

melabosch@hotmail.com

*Una vez que hemos catalogado y clasificado…*

***¿Cómo buscan los sistemas informáticos ?***

*Tipos de buscadores* → *bases de datos*

*Tipos de buscadores* → *buscadores Web*

# *El archivo y la búsqueda en bases de datos*

▪La mayor parte de los sistemas que acumulan información de manera persistente, incluídos muchos de los de referencias bibliográficas y bibliotecas digitales, se basan en datos estructurados con un modelo llamado de **entidad-relación:**

**Modelo de datos de entidad-relación o modelo relacional**

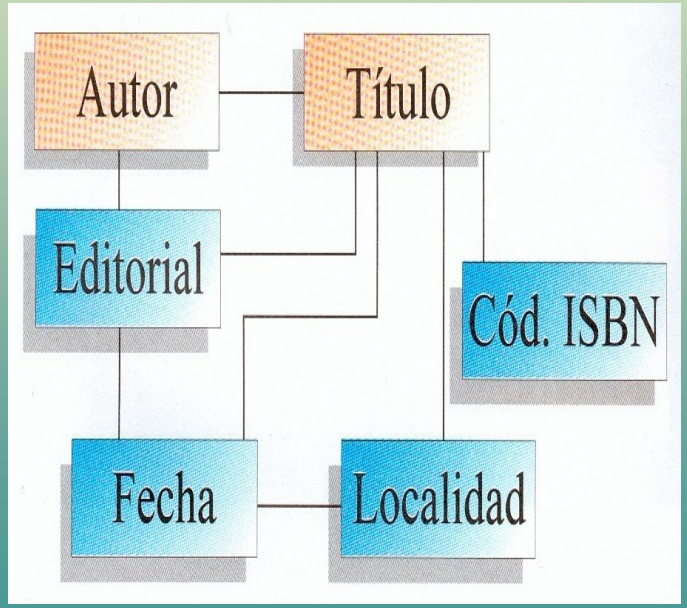
Origen: 1970, propuesto por Edgar Frank Codd, un investigador de IBM

revolución muy positiva pues constituyó un soporte lógico matemático fuerte y sencillo

**Antes de esa época:**

Archivos de registros y campos, al modo del original sistema ISIS, nacido a fines de los 60

**Es necesario mantener la estructura jerárquica en forma de árbol y cuando se necesita modificar una parte, toda la estructura se afecta**



# *El mercado de la búsqueda en bases de datos*

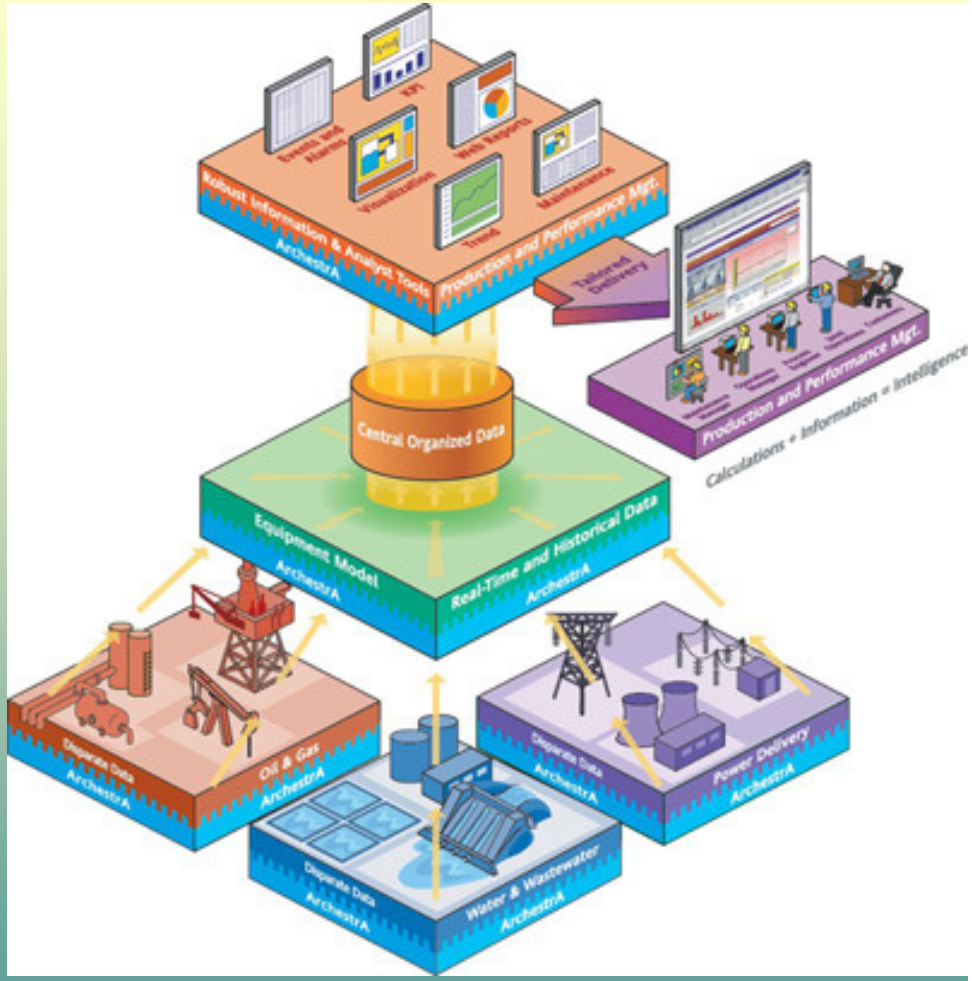
**Buscan en la bases de datos de documentación en las empresas:**

**Lideres del mercado de buscadores para bases de datos relacionales de las intranet de empresas:**

**Autonomy - Oracle - -Microsoft Index Server- Hyland / OnBase - Open Text- Ask Sam Web Publisher - Vertical Search Works-**

**Son costosos y están orientados a las intranet de las empresas, corporaciones, periódicos en línea.**

**También hay sistemas open source:mnGoSearch**




Uno de los líderes es Oracle que está en medio de una batalla judicial con Google por el uso de patentes para sistemas para tablets.

**Entonces el mundo de las intranet e internet confluye...**

*buscadores Web*

# 1. Motor de búsqueda textual: (*Text Search Engine*)

Basado en análisis léxico:

- Procesamiento: divide el texto en parágrafos, frases, palabras y también identifica entidades tales como direcciones de correo electrónico y direcciones Web, todos estos elementos para ser procesados son considerados como unidades acumulativas, técnicamente llamadas *tokens*, son sometidos a una serie de parámetros estadísticos con los que se establece un rango de enlaces, esta lista es la que se presenta como respuesta a nuestra pregunta.

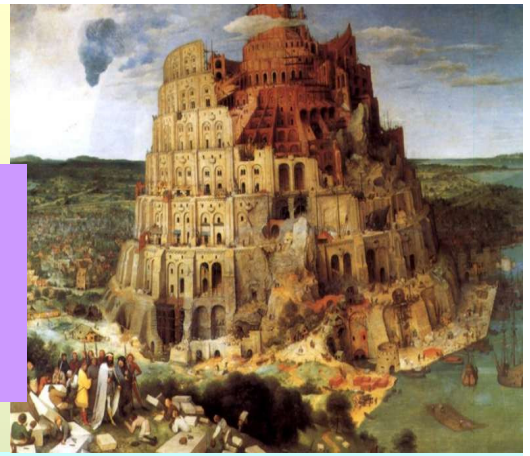
- Este tipo de motores son los primeros que aparecieron. Eran de este tipo el Gopher, creado en 1991 por Mark McCahill de la University of Minnesota y Veronica (**V**ery **E**asy **R**odent-**O**riented **N**et-wide **I**ndex to **C**omputerized **A**rchives)

## *2. Motores con indización de la semántica latente*
***(Latent semantic indexing, LSI)***


- Análisis del significado no literalmente explícito utilizando algoritmos con componentes estadísticos y léxicos.
- Latent semantic analysis (LSA) es una técnica de procesamiento del lenguaje natural (Natural Language Processing, NLP )
- Usa una base de datos de documentos para encontrar términos similares.
- En este tipo de motores se puede encontrar un cierto grado de sinonimia y devuelve los enlaces a los sitios Web que mejor se adaptan a nuestra búsqueda, el LSI no necesita tener el exacto término en una referencia para poder ofrecerlo como repuesta, puede usar aproximaciones de acuerdo con la estructura de sinónimos cuasi sinónimos que identifica.
- El motor Google utiliza este tipo de análisis, el componente estadístico es más fuerte que el de procesamiento de lenguaje y además tiene otros algoritmos como Page Rank

# 3. Motores de búsqueda semánticos,

***(Semantic Web search engines o Search engines of 3rd generation)***



- Intentan tomar el sentido de una palabra como factor para los algoritmos de ordenamiento y también pueden ofrecer al usuario posibilidades para desambiguar o refinar su consulta.
- Son llamados también motores de búsqueda de tercera generación, los cuales a su vez utilizan las otras dos tecnologías de búsqueda textual y de búsqueda de semántica latente a las que se suman otras específicas llamadas tecnologías de Web semántica.
- Estas son: ontologías, RDF (R*esource Description Forma*t) OWL (*Ontology Web Language*). Las tecnologías de Web semántica se basan en lógicas de descripción para dar cuenta de manera formal y computable de la semántica de los objetos de un sistema.

# 3. Motores de búsqueda semánticos, *(Semantic Web search engines Search engines of 3rd generation)*

**Tres tipos: a)**

**Buscadores semánticos orientados al usuario** *(User oriented Semantic Web search engines)*

- nos devuelven enlaces a páginas Web
- pueden usar internamente tanto tecnologías de Web semántica como de LSI
- Recuperación en diferentes formatos, imagen, sonido, etc para múltiples dispositivos
- Ya hay sistemas comerciales como Hakia, que busca incluso en los Twitters y E-Gains que también trabaja con redes sociales

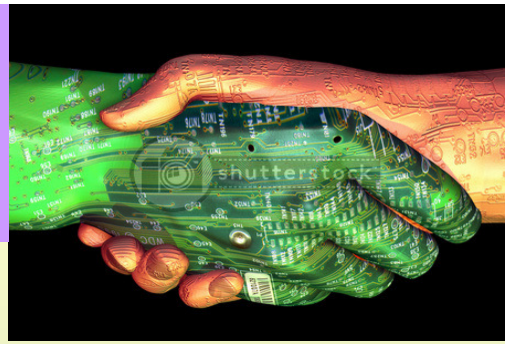
# 3. Motores de búsqueda semánticos, *(Semantic Web search engines Search engines of 3rd generation)*

b)

**Buscadores semánticos orientados a servicios Web** (*Semantic Web Services oriented)*


- No son para el usuario final
- Dan enlaces a útiles para los especialistas que trabajan en la construcción de recursos de Web semántica, devuelven enlaces a ontologías, archivos en OWL, a instancias de RDF
- Entre ellos tenemos: SOWL,WSE, Watson, Falcons, Sindice y Swoogle

# 3. Motores de búsqueda semánticos,
*(Semantic Web search engines Search engines of 3rd generation)*




**c)**

## Motores de búsqueda orientados a la Web social semántica *socio-semantic web (s2w)*

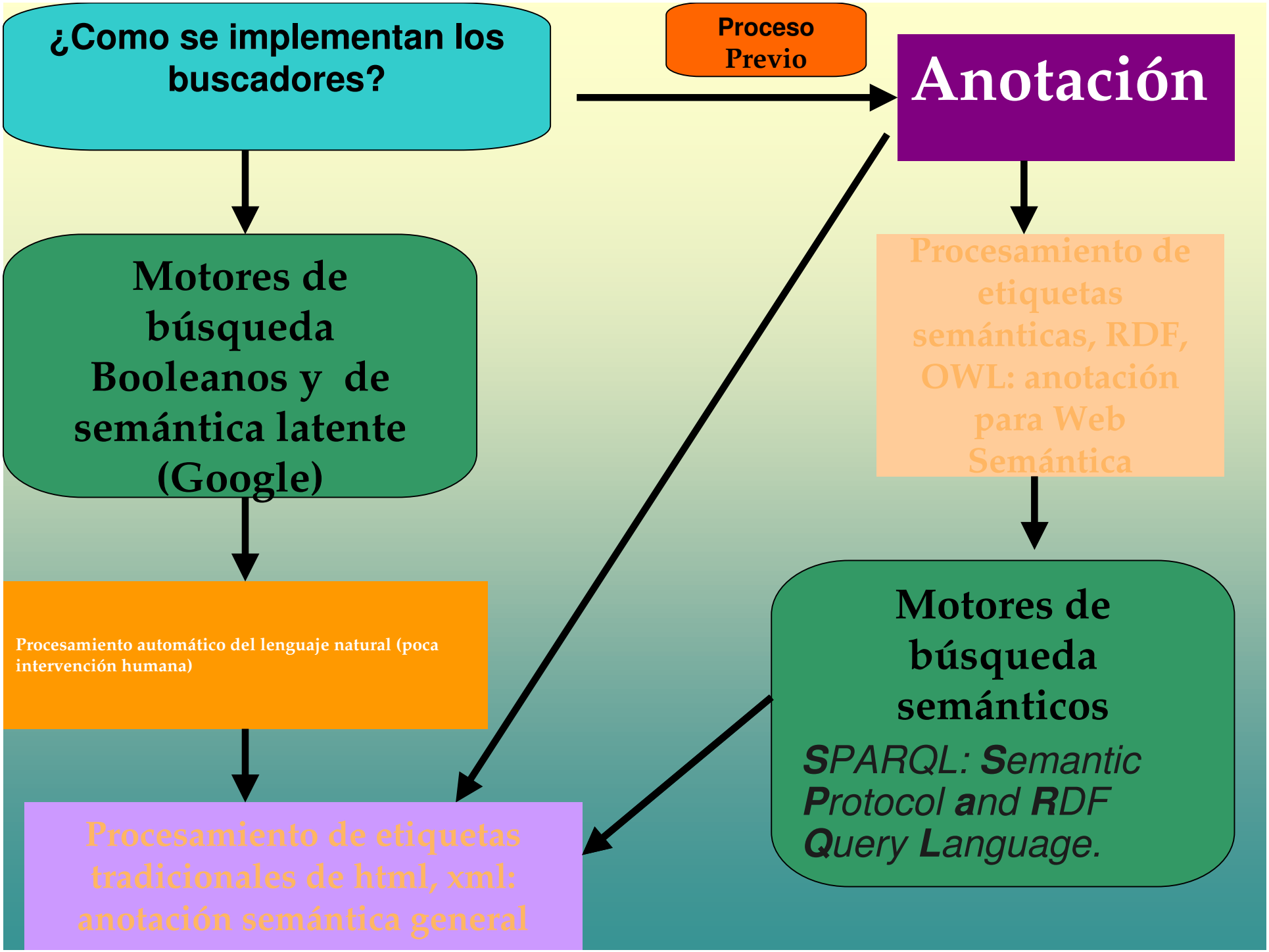
- Se proponen complementar la visión formal de la Web semántica con un acercamiento pragmático agregando a los lenguajes controlados creados con fuertes bases lógicas otros aspectos heurísticos basados en experiencias de prueba y error experimentadas por multitudes de usuarios que realizan etiquetado colaborativo (*folksonomy*)
- Entre ellos tenemos por ejemplo: http://www.stumpedia.com/
- La diferencia de este tipo de motores con los buscadores semánticos orientados al usuario es que utilizan microformatos de Web 2.0 (por ejemplo RSS) para poner etiquetas y usan para el trabajo cooperativo apoyado en computadoras Computer Supported Cooperative Work (CSCW). (ver: http://en.wikipedia.org/wiki/Social_Semantic_Web)

Buscadores de bases de datos,
Buscadores Web de varias generaciones:
**¿Qué hay en común en esta diversidad?**

1. **Anotación**

**Los instrumentos comunes a todos los Motores de búsqueda**

**2. Procesamiento del Lenguaje Natural**
**Natural Language Processing (NLP)**


¿Como se implementan los buscadores?
Proceso Previo
Anotación
Motores de búsqueda Booleanos y de semántica latente (Google)
Procesamiento de etiquetas semánticas, RDF, OWL: anotación para Web Semántica
Procesamiento automático del lenguaje natural (poca intervención humana)
Motores de búsqueda semánticos
SPARQL: Semantic Protocol and RDF Query Language.
Procesamiento de etiquetas tradicionales de html, xml: anotación semántica general

## Anotación semántica en general

Asociación de una entidad de datos con un elemento de tipo semántico que puede ser: esquema de clasificación, un tesauro, una nota, una glosa: larga tradición bibliotecaria y científica

**Anotaciones o etiquetado semántico, una antigua tradición con nueva instrumentación**


**Representar y organizar el conocimiento para transmitirlo y conservarlo**

**Anotación para Web semántica**

Objetivo hacer que las máquinas puedan comprender un dato en uno o varios sentidos y puedan usarlo para tomar decisiones y realizar acciones en determinadas y precisas situaciones

**Anotación proceso y resultado**

- ***El procedimiento para incluir datos comprensibles y procesables en la Web es agregar datos sobre los datos, es decir por medio de la creación de metadata.***
- ***Por eso es que usan también los términos de marcado semántico (semantic markup), etiquetado semántico (semantic tagging o semantic labelling***

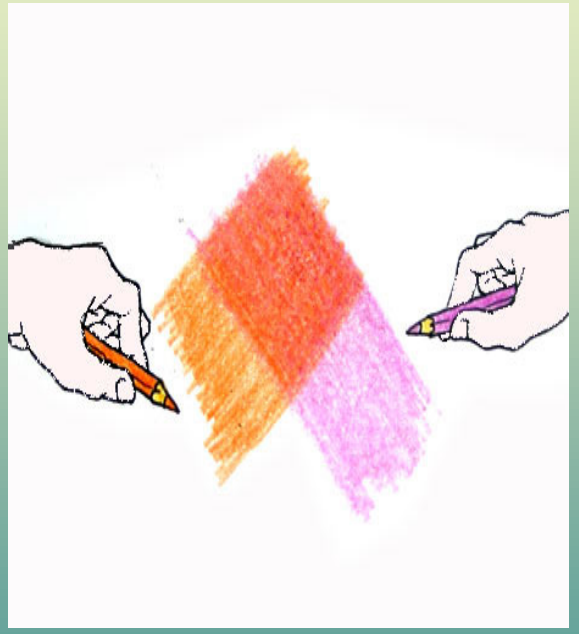
El concepto de anotación semántica (*semantic annotation*) conocido también como:

- marcado semántico (*semantic markup*)
- etiquetado semántico (*semantic tagging o semantic labelling)*
- El término **anotación** se ha ido imponiendo.

**Anotación proceso y resultado**

- Anotación en texto libre (*Free-text annotation*)

- **pueden ser cualquier tipo de comentarios, notas, explicaciones, referencias, sugerencias, correcciones**
- **cualquier tipo de indicación externa que puede ser agregada o incluida en un documento Web o en una parte de él**



**Anotación proceso y resultado**

- Anotación semántica en general

Significa la asociación de una entidad de datos con un elemento de tipo semántico que puede ser:

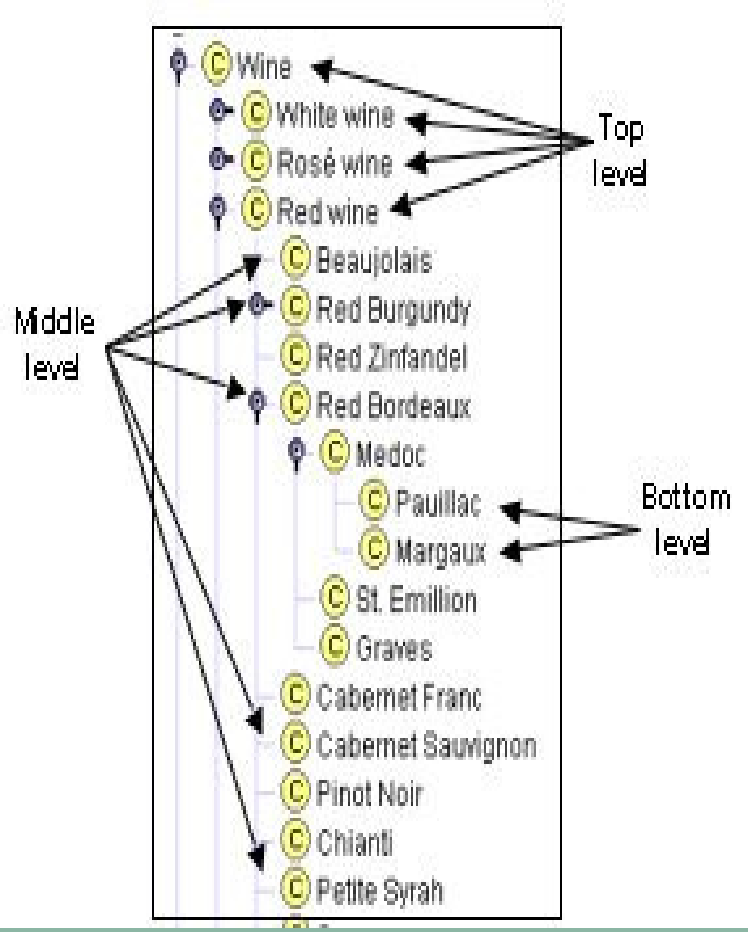
- Un esquema de clasificación, una ontología, un tesauro o cualquier otro instrumento de identificación de conocimiento en un repositorio de información.
- Por ejemplo las asignaciones de los descriptores de MeSH al las citaciones en MEDLINE, es decir una común indización por palabras claves o descriptores, hasta asignaciones de sentido mucho más complejas como los términos de la Gene Ontology a los productos genéticos en UniProt.

1a.

**Anotación para Web semántica: componentes**

**Una ontología que describe el dominio del sistema:**

- Términos de una ontología: términos controlados
- El objeto al que se refieren debe responder a una definición ontológica




**Ontologia en la Web semántica: conjunto de términos que refieren a objetos, términos expresados considerando sus propiedades y relaciones en un determinado dominio y sin que eso impida que en el desarrollo de una ontología intervengan diferentes enfoques metodológicos según el campo disciplinario de quienes aportan en la construcción de la ontología o en el dominio de aplicación de la misma**

**Anotación para Web semántica: componentes**

- Proceso de reconocimiento de la instancias de datos, es decir la detección de los objetos que responden a esa ontología, ya sea en un *corpus* de documentos o en un reservorio objetos multimediales que es manejado por el sistema que hace uso de la ontología.

**Anotación para Web semántica: componentes**

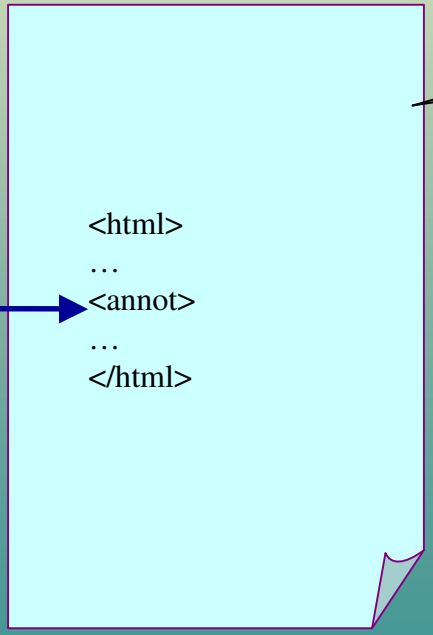
- El tercer componente es la acción que produce la anotación como resultado: se agrega contenido semántico a las instancias que responden a las propiedades y atributos de una ontología u otra estructura de representación de conocimiento
- puede ser manual, automática o semiautomática. Para ello se utilizan los herramientas de anotación semántica *(semantic annotation tools)*

# Tipos de herramientas herramientas de anotación semántica *(semantic annotation tools)*

## Anotación en línea *(Inline annotation):la metadata está incrustada en el documento*

Se centra en anotar usando RDF, OWL, es decir metadata que pueda ser interpretada por las computadoras

*Embedded metadata*




Se llama también:

Semantic Authoring

o

Bottom-up annotation

## Tipos de herramientas herramientas de anotación semántica (*semantic annotation* tools)

Standoff annotation: anotación separada, está escrita y archivada fuera del documento o de los objetos Web a los que refiere

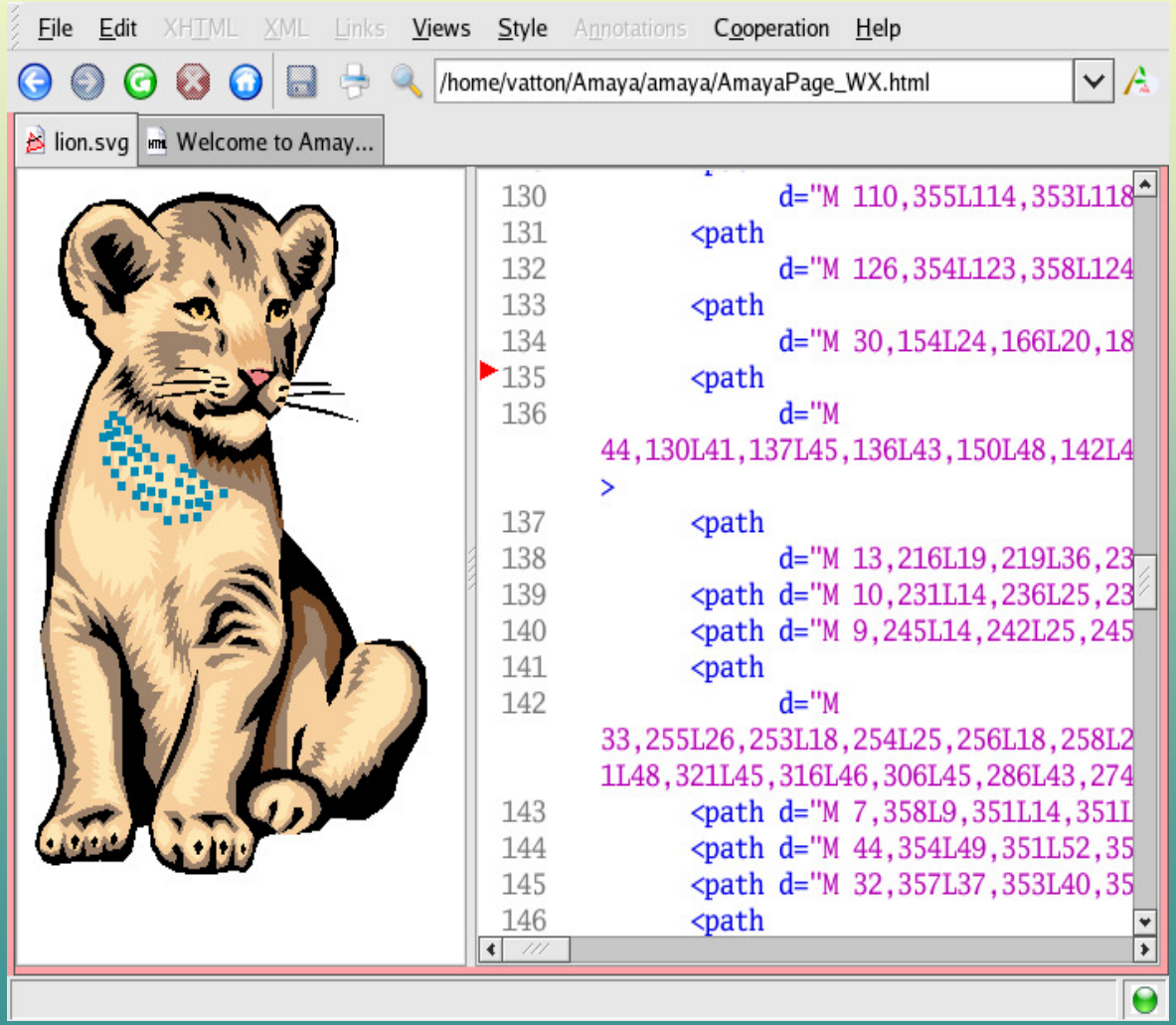

- Es preferible desde el punto de vista de la interoperabilidad

Llamada también anotación top-down

Valiosa para relevar información existente

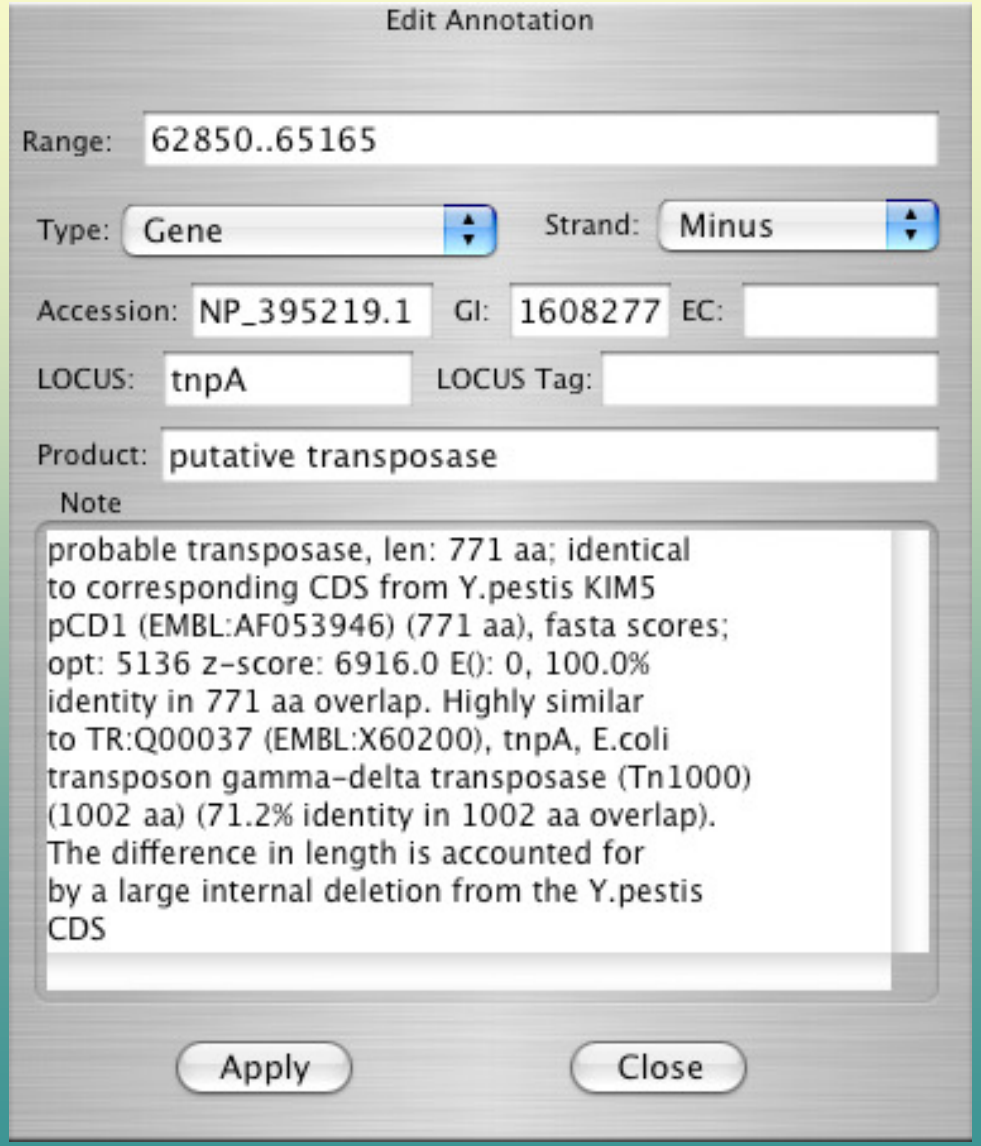
*Ejemplo de herramientas de anotación*

- **Amaya (embedded annotation)**
- **Fuente: http://www.w3.org/Amaya/screenshots/Overview.html.** [



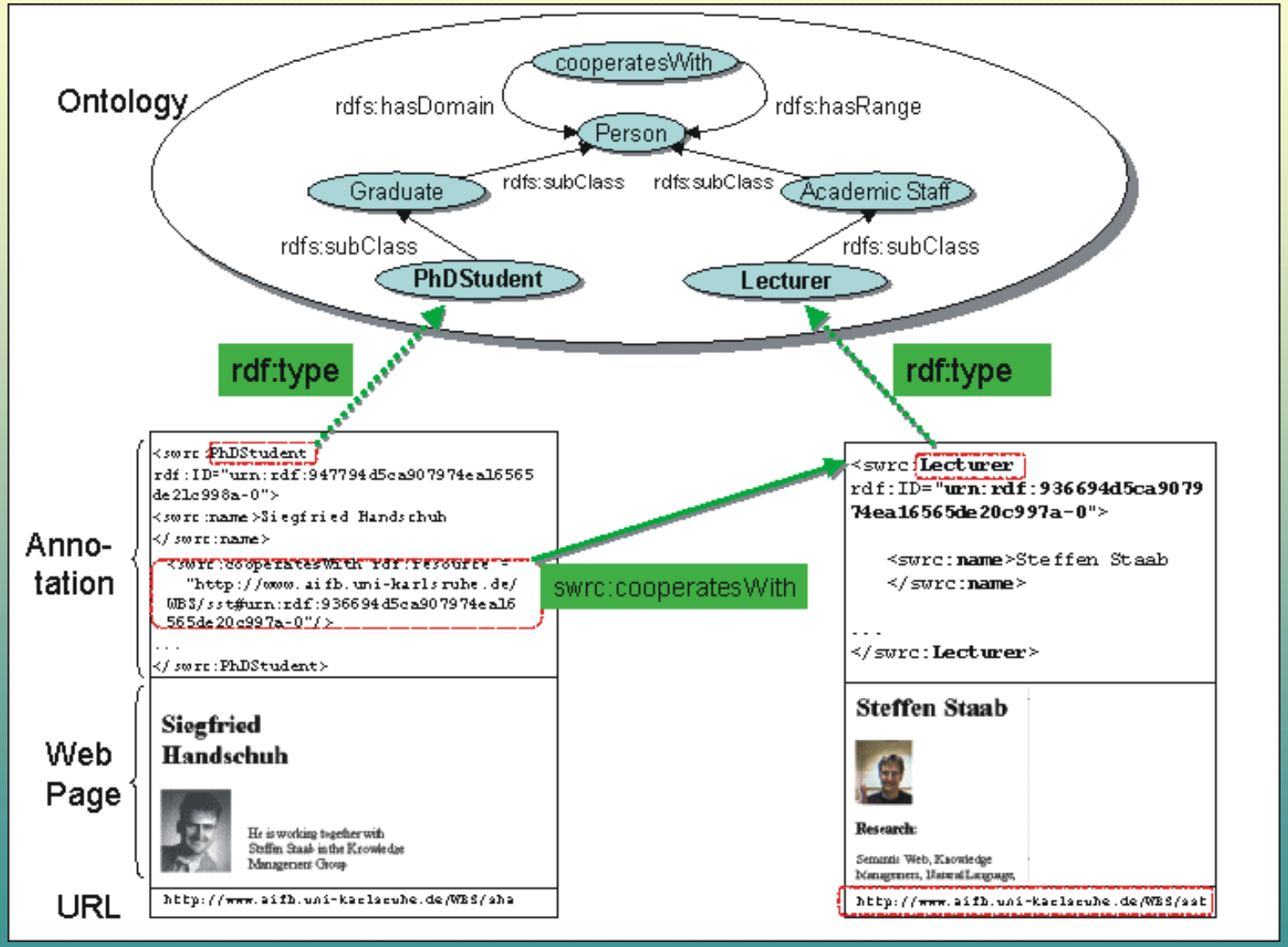
*Ejemplo de herramientas de anotación*

- **Gans, J.Multi-scale, Multi-genome, Multi-platform Visualization and Analysis. Los Alamos National Lab, Bioscience Division. http://public.lanl.gov/jgans/genomorama/genomorama_doc.html**



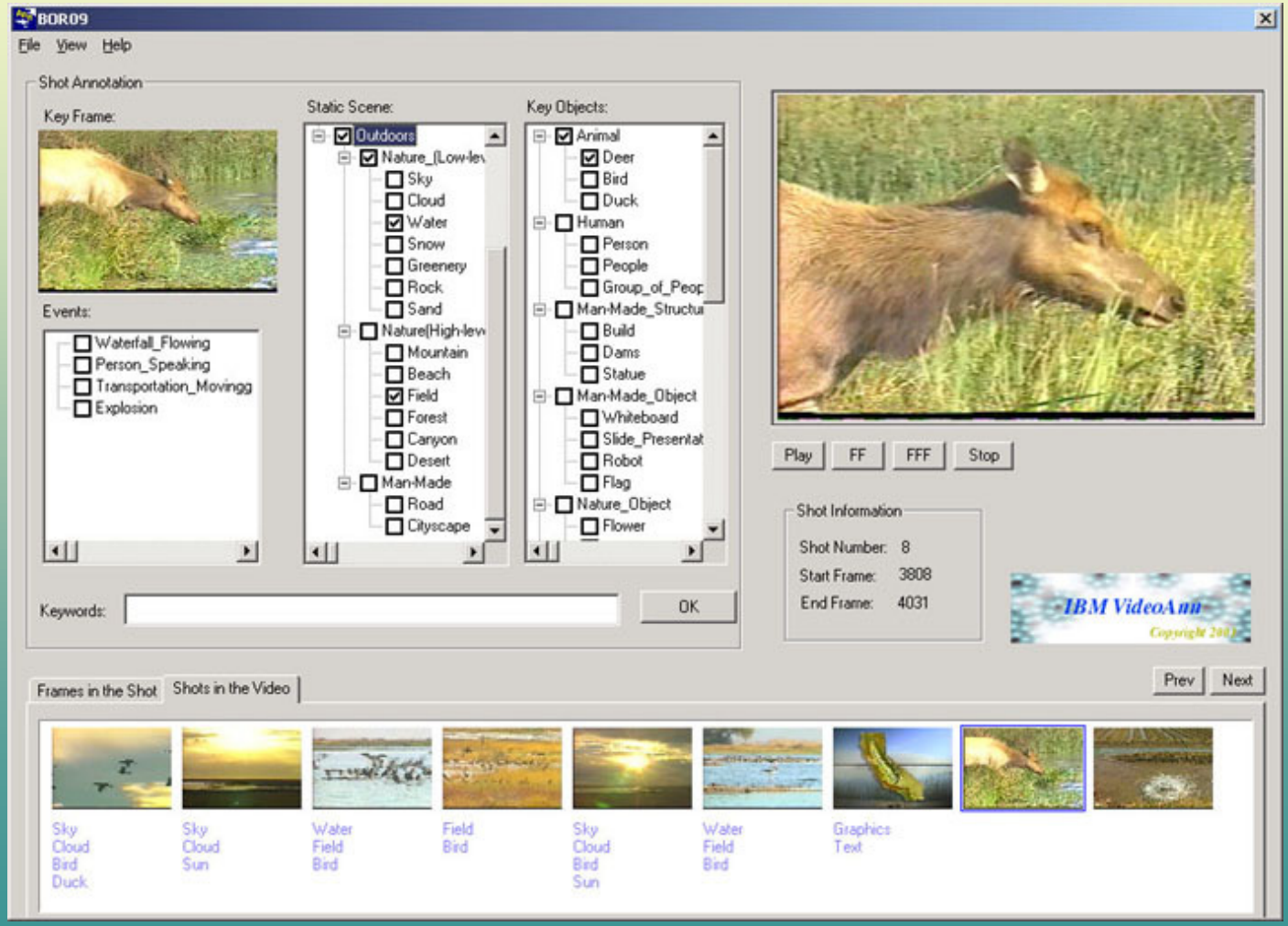
*Ejemplo de herramientas de anotación*

- **Anotación para Web Semántica. Staab, S. Handschuh, S. Authoring and Annotation of Web Pages in CREAM.** http://www2002.org/CDROM/refereed/506/ **[21-07-2009]**



*Ejemplo de herramientas de anotación*

- **Video Semantic Summarization Systems**
- **http://www.research.ibm.com/MediaStar/VideoAnn.html. [21-06-2009]**



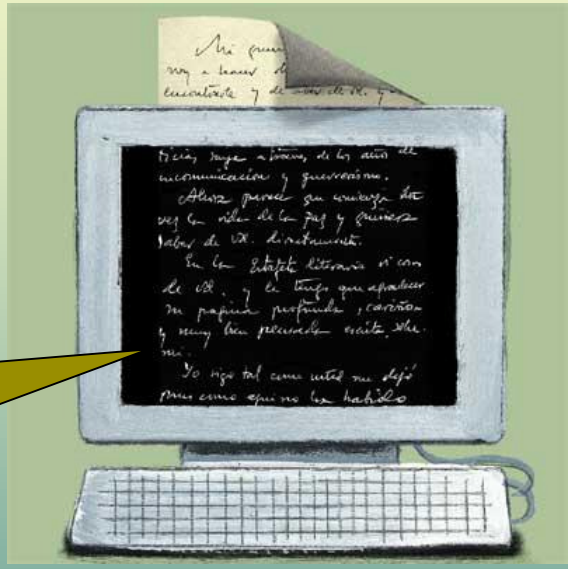
| Tool | Source | Type | Language | Origin |
|---|---|---|---|---|
| **Ontomat** | Html | Embedded | Formal (DAML+OIL) | Univ. of Karlsruhe, Germany |
| **Mnm** | | Embedded (file XML), Attached | Formal (DAML+OIL, RDF) | KMI The open Un. - Depart. of CS, Univ. of Sheffield. UK |
| | Html | | | |
| **Smore** | Photo, mail, html | Embedded | Formal (RDF, DAML+OIL, OWL) | University of Mariland |
| **Cohse** | Doc/ html | Attached | Formal DAML+OIL | Depart. of CS Univ of Manchester UK |
| **Trellis** | | Attached | Formal (OWL) | USC.Information Science Institute, Depart. of CS,Univ. of Sheffield. UK |
| **Melita** | Html | Attached | Formal | |
| **Kim** | Txt,Html, xml | Attacched | | Ontotext Lab, Sirma AI, Bulgaria |
| **Annotea** | Html/xml | Attached | socio-semantic web RDF/XML | W3C INRIA Rhône-Alpes W3C MIT/LCS |

# 2) El segundo componente de los **Semantic Web search engines**

- **Procesamiento del lenguaje natural- Natural Language Processing (NLP)**

**Al inicio:**

- proponía manipular palabras y partes de discurso con la idea de apoyar y desarrollar los estudios lingüísticos



***poderoso auxiliar para la investigación y la evaluación de los diferentes aspectos del lenguaje humano ya que permite estudiar grandes corpus de texto***

## Luego

- Más adelante en los años 60 la Inteligencia Artificial como disciplina se interesa en el procesamiento del lenguaje natural con propósitos prácticos como la traducción automática o la conversación de personas con máquinas.

# 2) El segundo componente de los **Semantic Web search engines**

## Procesamiento del lenguaje natural- Natural Language Processing (NLP)

Luego de la explosión de la Web el procesamiento de lenguaje natural es usado a la inversa para que las máquinas puedan comprender lo que las personas escriben o representan en los documentos, imágenes y videos

- *Esta inversión de dirección es la que ahora hace que algoritmos de procesamiento del lenguaje humano sean usados para extraer, desambiguar, etiquetar información en la Web.*

## Cierre: intercambio y puesta en común
## Mi experiencia con motores de búsqueda

- Alpha Wolfram **Orientado a informática, tiene aún mucho para desarrollar.**
- DuckDuckGo**: es un buscador muy bueno que no registra ni permite rastrear las búsquedas que realiza una persona. Ideal para mantener la confidencialidad.**
- Evri**: Ofrece noticias y recomienda, además, blogs, webs, tweets, citas, imágenes y videos. Su propuesta es el 'filtrado inteligente' . Está disponible en para móviles.**
- Hakia**: Los resultados se dividen en *Web, Noticias, Blogs, Twitter, imágenes, vídeo;* y pueden ser reclasificados por relevancia o fecha. Dependiendo del término, los resultados también pueden incluir un extracto de su entrada en Wikipedia. También se les etiqueta como 'creíbles' a los que provienen de fuentes de confianza.**
- Kngine**: utiliza a palabras clave, preguntas del usuario, y relaciones entre las palabras clave y enlaces afines. Los resultados los puedes compartir en Twitter, Facebook, Digg y Delicious. Hay quienes lo consideran mejor que**

## Cierre: intercambio y puesta en común
## Mi experiencia con motores de búsqueda

- **Kosmix: Muy dirigido a redes sociales para ver está pasando en la web de acuerdo a Yahoo Buzz, Digg, YouTube, Fark, Flickr y otras fuentes. También incluye las tendencias en Twitter.**
- **Powerset: Todos los resultados de búsqueda de Powerset provienen de Wikipedia, por lo que es la mejor manera de buscar en ese sitio, utilizando la semántica.**
- **Quora: muy valioso para hacer seguimientos de temas y mantenerse actualizado, incluso con Twitter.**
- **Truevert: Todos los resultados se filtran y se organiza desde una perspectiva relacionada a la conciencia ambiental.**