

Annotation

Oscar Corcho, Guadalupe Aguado Cea, Asunción Gómez-Pérez {ocorcho,lupe,asun}@fi.upm.es

Ontological Engineering Group Laboratorio de Inteligencia Artificial Facultad de Informática Universidad Politécnica de Madrid Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid, Spain

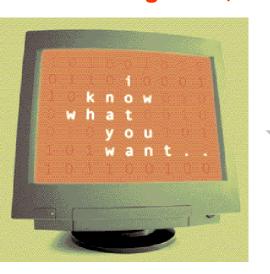
References

- Corcho O (2005) Ontology based document annotation: trends and open research problems. International Journal of Metadata, Semantics and Ontologies 1(1)
- Gómez-Pérez A, Fernández-López M, Corcho O (2004) Ontological Engineering. Springer-Verlag
- Handschuh S, Staab S (2003) Annotation for the Semantic Web. IOS Press



What is the Semantic Web?

"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. It is based on the idea of having data on the Web defined and linked such that it can be used for more effective discovery, automation, integration, and reuse across various applications."



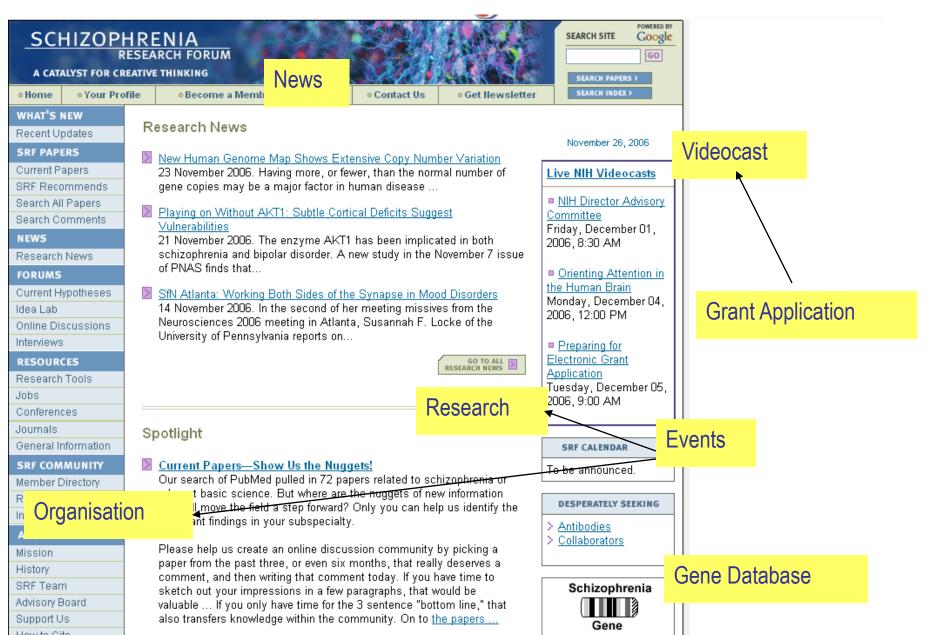
Onto ogies

Annotation

Hendler, J., Berners-Lee, T., and Miller, E. Integrating Applications on the Semantic Web, 2002, http://www.w3.org/2002/07/swint.html

Annotation assert facts using terms (metadata in RDF)

Represent terms and their relationships (ontology in RDFS/OWL)



Ontologies and Metadata (or Annotations)









4.2.1) Documentación administrativa:

- a) Original o fotocopia compulsada del documento acreditativo de la personalidad de la entidad solicitante.
- b) Poder notarial bastante del representante de dicha entidad, o documentación acreditativa de dicha cualidad.
- c) Fotocopia compulsada de la tarjeta de identificación fiscal de la Entidad.
- d) Original o copia con el carácter de auténtica o lotocopia compulsada de los Estatutos debidamente legana dec.
- Original o fotocopia compulsada de la siguiente documeniación acreditativa del cumplimiento de Obligaciones Tributarias y de Seguridad Social;

Recipo del año anterior a la cor económicas, o en su caso, exenci

Certificaciones administrativa por las Administraciones correspnistración Tributaria y de la Teso de conformidad con lo estableci Presupuestaria y en su caso, exey del Impuesto de Sociedades.

```
Tanding_opportu: DocumentationItem
   rdf:about="http://protece.stanford.edu/tmp_namespace#tmp_Instance_10032" rdfs:label="BOE">
   <tmp_namespace:Description>Fotocopia compulsada de la tarjeta de identificación fiscal de la
     Entidad.</tmp_namespace: Description>
   Funding_Opportu:isNeededToApplyFor
     df:resource="http://protege.stanford.edu/tmp_namespace#tmp_Instance_10004" />
  </Funding Opportu:DocumentationItem>
- <Funding_Opportu:DocumentationItem
   rdf:about="http://protege.stanford.edu/tmp_namespace#tmp_Instance_10033" rdfs:label="BOE">
   <tmp_namespace:Description>Original o copia con el carácter de auténtica o fotocopia compulsada de
     los Estatutos debidamente legalizados.</tmp_namespace:Description>
   <Funding Opportu:isNeededToApplyFor</p>
     rdf:resource "http://protege.stanford.edu/tmp_namespace#tmp_Instance_10004" />
 </Funding_Opportu: DocumentationItem>
- <Funding Opportu: DocumentationItem
   rdf:about="http://proteqe.stanford.edu/tmp_namespace#tmp_Instance_10034" rdfs:label="BOE">
   <tmp_namespace:Description>Original o fotocopia compulsada de la siguiente documentación
     acreditativa del cumplimiento de Obligaciones Tributarias y de Seguridad Social: Recibo del año
     anterior a la convocatoria del Impuesto sobre actividades económicas, o en su caso, exención
     concedida por el órgano competente. Certificaciones administrativas con el carácter de positivas
     expedidas por las Administraciones correspondientes de la Agencia Estatal de Administración
     Tributaria y de la Tesorería Territorial de la Seguridad Social, de conformidad con lo establecido en
     el artículo 81 de la Ley General Presupuestaria y en su caso, exenciones del Impuesto del Valor
```

Añadido y del Impuesto de Sociedades.</tmp_namespace:Description>

Anotación de contenidos Web. Grados de detalle

Distintos tipos de anotación según el vocabulario utilizado

Basada en Dublin Core

The *contributor* and *creator* is the flight booking service "www.flightbookings.com".

The *date* would be January 1st, 2003, in case that the HTML page has been generated on that specific date.

The *description* would be something like "flight details for a travel between Madrid and Seattle via Chicago on February 8th, 2004".

The document format is "HTML".

The document language is "en", which stands for English

Flight details

Outbound

Leaving from **Madrid** - Barajas - Spain on Saturday 08 February 2003 at **11:50** Arriving in **Chicago** - O'Hare International - United States of America same day at **14:10** Airline: American Airlines Flight No. AA 7615 Type of aircraft: Airbus Industrie A340 All Series PAX/H

Leaving from **Chicago** - O'Hare International - United States of America on Saturday 08 February 2003 at **16:48**Arriving in **Seattle** - Seattle/Tacoma International - United States of America same day at **19:23**Airline: American Airlines

Flight No. AA 1605
Type of aircraft: non referenced/B

Basada en tesauros

refers to the kingdom of Spain in Europe.

Madrid is a reference to the term with ID 7010413 in the thesaurus, which refers to the city of Madrid in Spain. Spain is a reference to the term with ID 1000095, which

Chicago is a reference to the term with ID 7013596, which refers to the city of Chicago in Illinois, US.

United States of America is a reference to the term "United States" with ID 7012149, which refers to the US nation. Seattle is a reference to the term with ID 7014494, which refers to the city of Seattle in Washington, US.

Basada en ontologías

Concept instances relate a part of the document to one or several concepts in an ontology. For example, "Flight details" may represent an instance of the concept Flight, and can be named as AA7615_Feb08_2003, although concept instances do not necessarily have a name.

Attribute values relate a concept instance with part of the document, which is the value of one of its attributes. For example, "American Airlines" can be the value of the attribute companyName.

Relation instances that relate two concept instances by some domain-specific relation. For example, the flight AA7615 Feb08 2003 and the location Madrid can be connected by the relation departurePlace

Anotación de contenidos Web. Dimensiones

- Fuentes
 - Tipos de fuentes
 - Texto: HTML, XML, PDF, etc.
 - Multimedia: imágenes, video, audio, etc.
 - Servicios Web
 - Origen de la fuente de datos
 - Estático: ficheros
 - Dinámico: bases de datos y formularios
- Tecnologías utilizadas
 - Knowledge extraction
 - NLP, IE, Layout
 - Wrapper generation
 - Toolkits, ML, Browsing

- Proceso de anotación
 - Mantenimiento (adaptabilidad a los cambios en la fuente)
 - Verificación
 - Robustez
 - Auto-adaptabilidad
 - Supervisión de las anotaciones
 - Manual
 - Supervisado (semiautomático)
 - No supervisado (automático)
- Grado de formalidad
 - Etiquetas (Web 2.0)
 - Basada en ontologías

Different types of sources















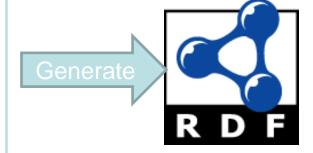






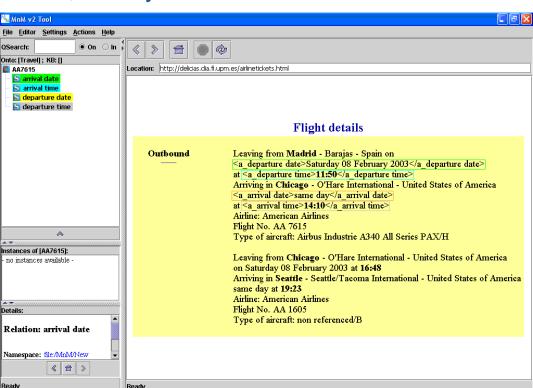


Different sources



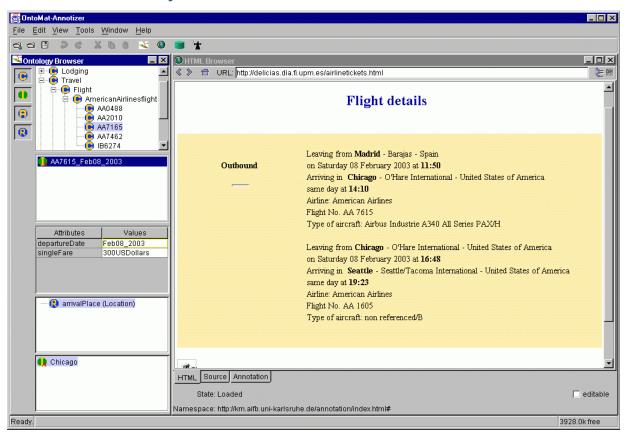
Herramientas de anotación. MnM

- Aplicación standalone que integra un navegador Web y un visor de ontologías
- Anotación manual con drag&drop
- Anotación semi-automática y automática utilizando herramientas de extracción de información (Amilcare)
- Anotaciones generadas en OCML, RDF y XML



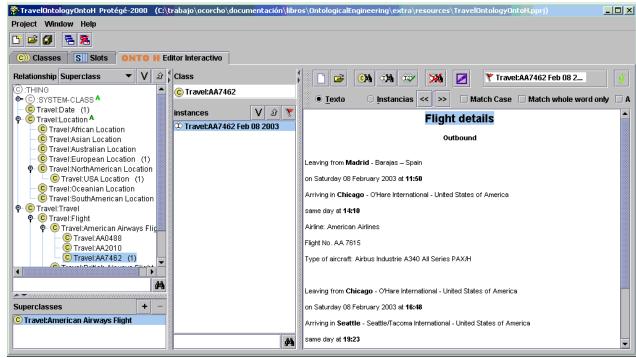
Herramientas de anotación. OntoMat

- Aplicación standalone que integra un navegador Web y un visor de ontologías
- Anotación manual con drag&drop
- Anotaciones generadas en RDF y OWL



Herramientas de anotación. ONTO-H

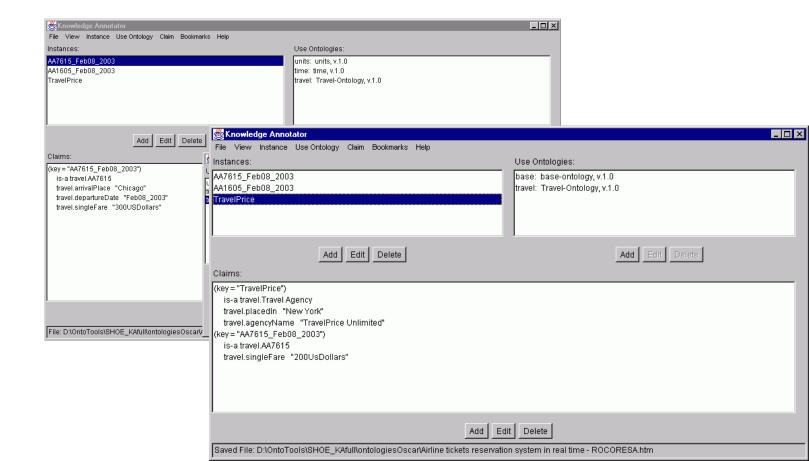
- Plug-in de Protégé que integra un visor de documentos RTF y el visor de ontologías por defecto
- Anotación manual con drag&drop
- Anotación en cascada de conceptos relacionados, mediante reglas de anotación declarativas
- Anotaciones generadas en el formato de Protégé, exportable a diversos lenguajes



Herramientas de anotación. SHOE Knowledge Annotator

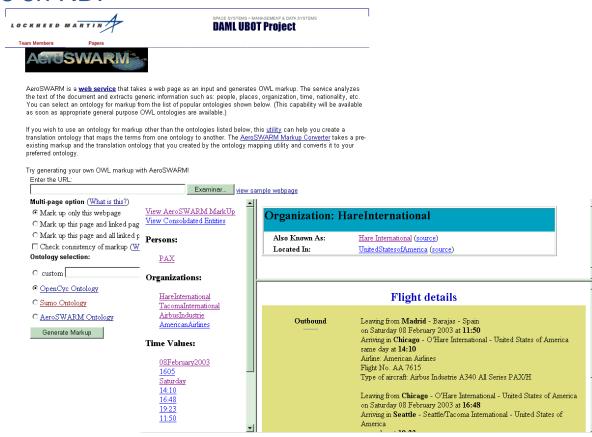


- Aplicación standalone que no contiene navegador Web
- Anotación manual
- Anotaciones generadas en SHOE



Herramientas de anotación. AeroSWARM

- Servidor Web de anotación de documentos Web disponibles en una URL
- Anotación automática, basada en un conjunto de ontologías predefinidas: OpenCyc, SUMO y AeroSWARM
- Anotaciones generadas en RDF



Herramientas de anotación. Knowledge Parser



- Conocimiento de wrapping explícito
- Bootstrapping

Estrategias

- Heurística
- Backtracking
- Fuerza bruta, etc.

Diferentes operadores

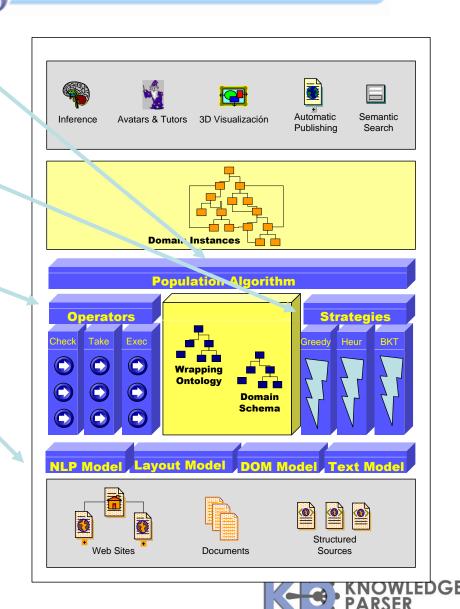
- In-Row
- Is-Proper-Name
- Is-Integer-Greater-Than, etc.

Diferentes pre-procesados

- Lenguaje natural
- Layout
- XML/DOM
- Texto plano

Sectores en los que se ha aplicado

- Cultural
- Subvenciones
- Financiero
- Relaciones internacionales



Anotación de contenidos Web. Dimensiones

- Fuentes
 - Tipos de fuentes
 - Texto: HTML, XML, PDF, etc.
 - Multimedia: imágenes, video, audio, etc.
 - Servicios Web
 - Origen de la fuente de datos
 - Estático: ficheros
 - Dinámico: bases de datos y formularios
- Tecnologías utilizadas
 - Knowledge extraction
 - NLP, IE, Layout
 - Wrapper generation
 - Toolkits, ML, Browsing

- Proceso de anotación
 - Mantenimiento (adaptabilidad a los cambios en la fuente)
 - Verificación
 - Robustez
 - Auto-adaptabilidad
 - Supervisión de las anotaciones
 - Manual
 - Supervisado (semiautomático)
 - No supervisado (automático)
- Grado de formalidad
 - Etiquetas (Web 2.0)
 - Basada en ontologías

Social tagging/annotation

- ¿Qué es la Web2.0?
 - http://es.youtube.com/watch?v=nsa5ZTRJQ5w
 - http://es.youtube.com/watch?v=PL-ywltLjzk
- ¿Qué es el "etiquetado social semántico?
 - Poner, gestionar etiquetas –metadatos- de forma colaborativa en la Web para clasificar el contenido (texto, video, fotos, música, etc..)
- Finalidad:
 - Clasificar páginas web según utilidad, facilidad de uso, adecuación, etc.
 - Mejorar y adaptar los contenidos de una página web de modo similar a un wiki
- Utilidad:
 - Como herramienta colaborativa, para tratar sobre la adecuación de los contenidos de un recurso
 - Como forma de visibilidad de los usuarios en la web: defender o crtiicar determinados temas
- Términos utilizados en inglés
 - Social bookmarking, collaborative tagging, folksonomy, social classification, social indexing, folksonomies

Social Tagging/Annotation systems



Del.icio.us

- Etiquetado hecho por creadores de contenidos o usuarios no por expertos
 - Las etiquetas son descriptores de una palabra
 - Se utilizan para describir metadatos del bookmark o calificarlo: * OK, *** Very good, **** Outstanding
 - El usuario puede asignar las que quiera y cuantas quiera
- Añadir etiquetas resulta más fácil y más flexible que encajar la información en carpetas o categorías ya establecidas.
 - Ejemplo: si se quiere guardar un artículo sobre cómo hacer un pastel, se puede poner recipes sweets yogurt o cualquier otra etiqueta que resulte sugerente

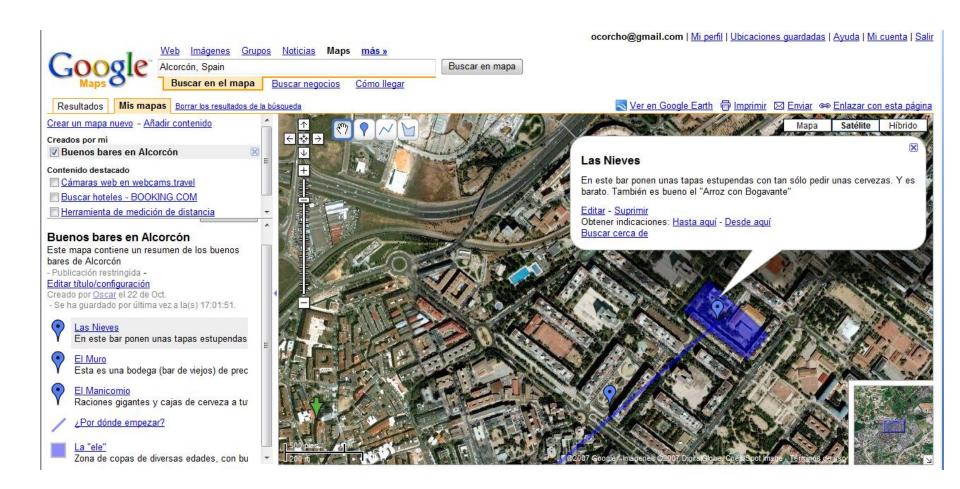
Annotation in del.icio.us

del.icio.us	
uri	http://www.foodsubs.com/
description	The Cook's Thesaurus
notes	pictures of ingredients and substitutions
tags	food reference
save	
recommended tags export food guide search	
your netwo for:joshua fo	
popular tac	reference recipes thesaurus Dictionary cook

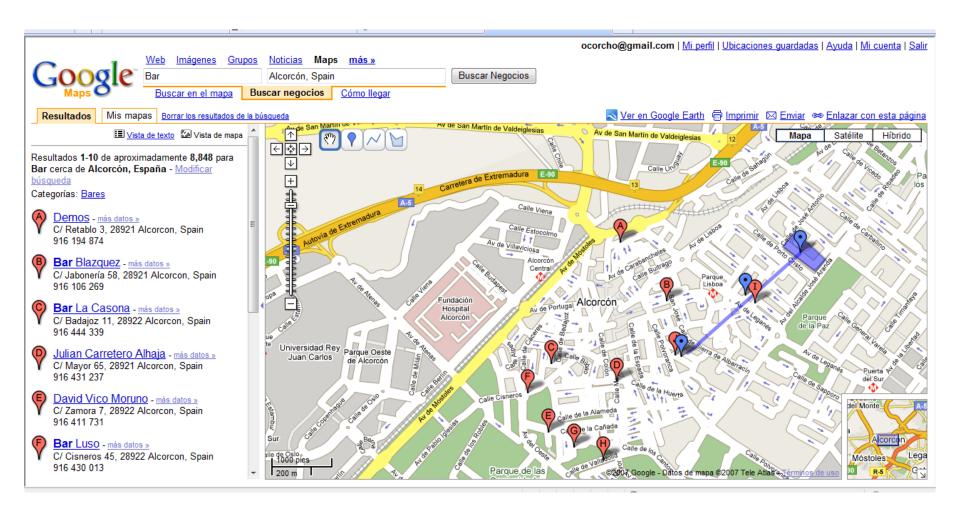
Del.icio.us

- Se pueden ver
 - las etiquetas (bookmarks) de otras personas sobre un tema.
 - las etiquetas más populares sobre ese tema
- Organización
 - Posibilidad de crear clusters para sacar inferencias
 - Universo o nube de etiquetas: (tag cloud) que se pueden compartir o restringir.
 - Se pueden ver los enlaces que traten del mismo tema
- Indicadores de actualización de etiquetas y novedades (web feed)
- Posibilidad de incluir estadísticas

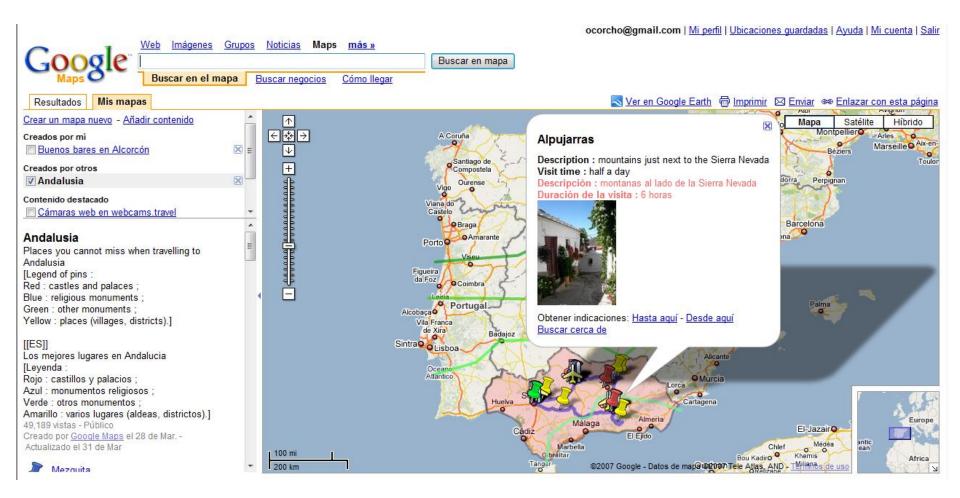
Google Maps tags



Google Maps tags



Google Maps tags



Folksonomies

- Origen: folks: gente (colegas) taxonomy (taxonomias)
- Los internautas aportan sus propias etiquetas (tags) e interpretan las de los demás, sin un criterio especificado
- Las relaciones que se crean no son establecidas:
 - Parte del "encanto" está en el carácter subversivo de seleccionar las propias etiquetas.
 - Contradicción con los motores de búsqueda que utilizan etiquetas aceptadas por una comunidad
- El etiquetado social no forma parte de los protocolos web, sino que se decide a nivel del sitio web.
- ¿Qué se etiqueta?
 - TODO!!!! Fotografías, canciones, texto, referencias bibliográficas, enlaces de una universidad, sitios web, entradas de blogs, etc. etc.

A sample folksonomy: Panoramio

(f) Inicio Mapamundi Lugares Etiquetas Subir fotos Registrarse Entrar Foro Blog Ayuda

Etiquetas populares

Buscar etiqueta

1 2003 2004 2005 2006 2007 amsterdam animals argentina australia austria barcelona beach belgium berlin brasil brazil bridge budapest building buildings california canada castle chile china church city croatia denmark deutschland egypt england españa finland florida flowers france germany greece hamburg hotel hungary india ireland italia italy japan kirche lake landscape lighthouse london madrid mallorca mexico moscow mountain mountains nature netherlands new york new zealand night norway panorama paris park poland portugal prague river roma rome russia san francisco scotland sea snow Spain sunrise sunset sweden switzerland thailand turkey usa water waterfall winter

 Go

Problems

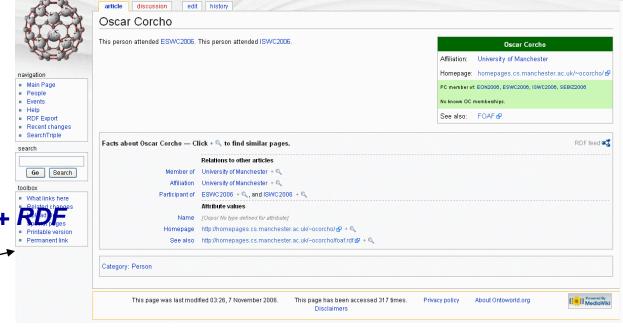
- Idiosyncratic folksonomic classification, although considered beneficial by some, is viewed by others as a distinct limitation.
- Lack of terminological precision produces inconsistent and unreliable results:
 - synonymy: polución, contaminación, solicitud, aplicación
 - homonymy: comando (una orden en programa) comando (de guerra)
- Lack of stemming (meta-noise)
 Heterogeneity of users and contexts. inaccurate or irrelevant tags
- Lack of a hierarchical structure for the tagging system makes the terms relevant to what people are describing, BUT fails to show their relevancy or relationship to other objects of the same type.
- SO metadata tags need to be defined in a formal way at the time of scripting or programming.

Anotación de contenidos Web. Dimensiones

- Fuentes
 - Tipos de fuentes
 - Texto: HTML, XML, PDF, etc.
 - Multimedia: imágenes, video, audio, etc.
 - Servicios Web
 - Origen de la fuente de datos
 - Estático: ficheros
 - Dinámico: bases de datos y formularios
- Tecnologías utilizadas
 - Knowledge extraction
 - NLP, IE, Layout
 - Wrapper generation
 - Toolkits, ML, Browsing

- Proceso de anotación
 - Mantenimiento (adaptabilidad a los cambios en la fuente)
 - Verificación
 - Robustez
 - Auto-adaptabilidad
 - Supervisión de las anotaciones
 - Manual
 - Supervisado (semiautomático)
 - No supervisado (automático)
- Grado de formalidad
 - Etiquetas (Web 2.0)
 - Basada en ontologías

Wikis semánticos



Presentación de RDF
Rhizome

Edición de páginas HTML + RDE

IkeWiki

SemanticMediaWikit

SemPerWiki

SweetWiki

WikSAR

Inclusión de instancias de ontologías

OntoWiki

COW

 \longrightarrow

SemanticWikipedia

👗 Log in / create account

Anotación atributo-valor no basada en ontologías

DiamondWiki

Escritorios semánticos (incluyendo e-mail semántico)

