Datos abiertos enlazados: situación actual y perspectivas

Christian Sifaqui

II Congreso de Bibliotecas Universitarias y Especializadas

4 de junio de 2015

Motivación 1

Mundo pre-coordinado y mundo post-coordinado

- “Falsa” tensión
- Bibliotecología - Ciencias de la Computación

Ejemplo portal de noticias

Christian Sifaqui, “Gestión digital de información de prensa”, Serie Bibliotecología y Gestión de Información, 2014, n. 92. http://eprints.rclis.org/24155/

Motivación 1

Pre-coordinado

Nombre de la publicación:

"EL SIGLO"

Ciudad SANTIAGO

Fecha: Año 1973 Mes Ante Día 20

Página 9 Columna 4

Ubicación del recorte F 21 e

Biblioteca del Congreso Nacional — Anex

Regresaron dirigentes del PR desde Argentina

Regresaron al país los dirigentes del Partido Radical, encabezados por el senador Anselmo Sule, quienes permanecieron en Buenos Aires por espacio de una semana, celebrando conversaciones con el líder de la Unión Cívica Radical de Argentina, Ricardo Balbín.

Los dirigentes radicales manifestaron que invitaron a Ricardo Balbín a nuestro país, y que éste expresó que lo haría en los primeros días de noviembre próximo.

La delegación que viajó a la hermana república la integraron además Carlos Morales, miembro de la Comisión Política del Partido Radical; Mario Hurtado, dirigente nacional y Alejandro Montecinos de la Juventud Radical Revolucionaria.

Asimismo, señalaron que el próximo martes ofrecerán una conferencia de prensa en la que darán cuenta de su visita a la Argentina y de las varias reuniones celebradas con Balbín y además se referirán el momento político que vive el país.

Motivación 1

Post-coordinado

24 Oct. 2012 - 02:09 pm

24 Oct. 2012 - 02:10 pm

24 Oct. 2012 - 02:10 pm

Inicio > Noticias

CHRISTIAN SIFAQUI

1

VOLVER AL LISTADO

Imprimir Enviar por correo Guardar en su cuenta Crear zip Crear pdf

EDITAR

Serán enviados a planta certificadora del Ministerio de Transportes:

CARABINEROS ORDENA REVISAR ALCOTESTS QUE HAN ENTREGADO RESULTADOS ABULTADOS

El Mercurio 24 Octubre 2012 Cuerpo C p. 7 Nacional
(32.5 x 29.0 cm., aprox. 942.5 cm² - 50.7% de 1 página)

Titulares de la Corte de Apelaciones y de la Asociación de Magistrados insistieron ayer en sus dudas por las discrepancias en registros de alcoholemias. Gobierno afirma que examen espirométrico es sólo referencial.

Las dudas en torno a la precisión de los resultados que arrojan los aparatos de alcotest se acrecentaron luego de que trascendiera que la alcoholemia practicada a la actriz Daniela Ramírez habría arrojado 1,6 gr/l de sangre, es decir menos de la mitad de los 3,39 gr/l que marcó la joven al ser controlada por Carabineros la semana pasada en el test espirométrico.

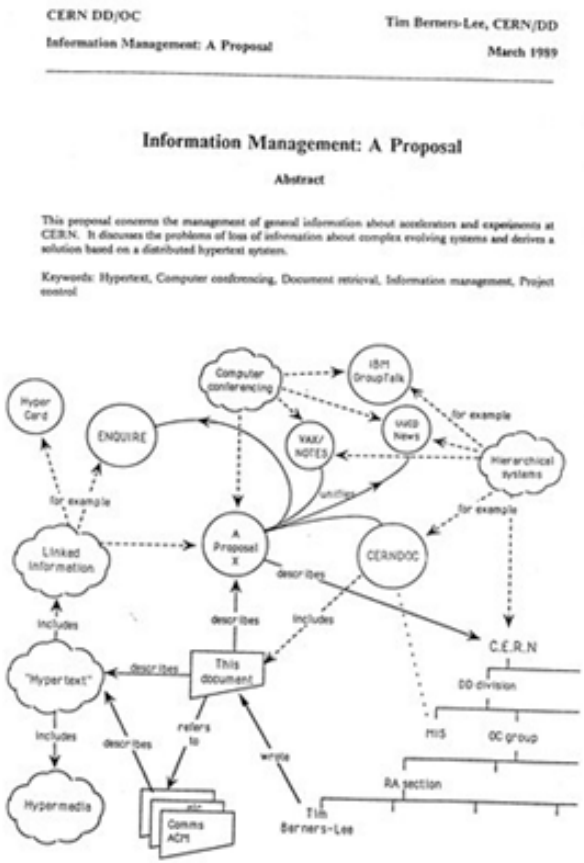
El dato surgió luego que el lunes, en la audiencia de formalización del ex futbolista Manuel Neira, el juez Patricio Souza planteara que existen dudas entre los jueces debido a los altos registros alcohólicos obtenidos en el último tiempo a través de los alcotest. En este escenario, Carabineros decidió ayer enviar al Centro de Certificación y Control Vehicular del Ministerio de Transportes (denominada 3CV) los aparatos que han arrojado resultados dudosos, según fuentes de la policía uniformada. La medida apunta a detectar si existió alguna anomalía en su funcionamiento, pese a que la institución insistió en que han sido sometidos a certificación recientemente.

En paralelo, tanto la Corte de Apelaciones como la Asociación de Magistrados respaldaron ayer a Souza, e incluso se advirtió que la diferencia entre ambas mediciones podría llevar a error en las imputaciones que se hacen a una persona que fue sorprendida conduciendo con alcohol.

Motivación 2

La web fue creada para compartir “documentos”

Tim Berners-Lee, “Information Management: A proposal”, 1



CERN DD/OC

Tim Berners-Lee, CERN/DD

Information Management: A Proposal

March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext system.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Motivación 3

El valor de una red es la “conexión”

- Metcalfe's Law
- Reed's Law

David Reed, “The Law of the Pack”, Harvard Business Review, February 2001, 23-24

James Hendler and Jennifer Golbeck, “Metcalfe's Law, Web 2.0, and the Semantic Web”, Journal of Web Semantics 6(1): 14-20, 2008

Motivación 4

¿Cómo clasificar un documento en mi computador?

- Documento ≠ archivo computacional
- Herramientas: jerarquía rígida de directorios y nombres mnemotécnicos para los archivos computacionales

Deborah Barreau and Bonnie Nardi, "Finding and Reminding: File Organization from the Desktop", SIGCHI Bulletin, 27(3), July 1995 (buscar en directorios, en vez de usar search)

Scott Fertig, Eric Freeman and David Gelernter, "Finding and Reminding Reconsidered", SIGCHI Bulletin, 28(1), January 1996 (es porque no hay otras opciones)

- Énfasis en dónde está, en vez de qué es el documento

Bolsa de documentos → Google desktop (septiembre 2011 dejó de actualizarse)

Motivación 4

Clasificar documentos en la web

- Yahoo Directory (cerrado en diciembre 2014)
- www.dmoz.org

navegar

El poder de la web (encontrar lo que se necesita) se produce a través del espacio de enlaces que emerge de las páginas web. Por ejemplo, algoritmo PageRank

- Google

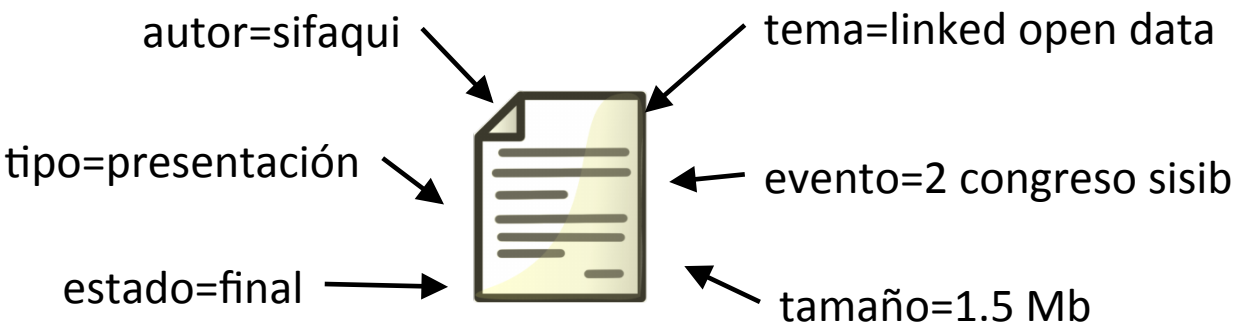
buscar

Sergey Brin, Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine", Proceedings of the seventh international World Wide Web Conference, 1998

Motivación 4

Documento trata de cosas, entidades, etc.

Disponer “atributos” con semántica



Motivación 4

- Marcadores sociales, folcsonomía, tagging
- Taxonomías
- Ontologías

- http://www.shirky.com/writings/ontology_overrated.html (2005) pero Folcsonomía falla fuera de los ámbitos sociales

Motivación 5

Datos abiertos (open data)

es un idea que impulsa la publicación de datos de forma libre y asequible a cualquier persona, para que sean usados y republicados sin restricciones de ningún tipo

World Wide Web

Repleta de información

Orientada al ser humano

- para comprender el contenido de una página
- para relacionar contenidos dentro de una página (textos, imágenes, videos, etc.)

World Wide Web

Buscadores actuales funcionan bien, pero orientados al keyword

- análisis de palabras y textos
- análisis de los enlaces

¿y consultas más complejas, con "semántica"?

Ejemplo: "diputados o senadore cuyos ermanos hayan sido jueces"

World Wide Web

Lo que “entiende” un programa

Lo que “entiende” un ser humano

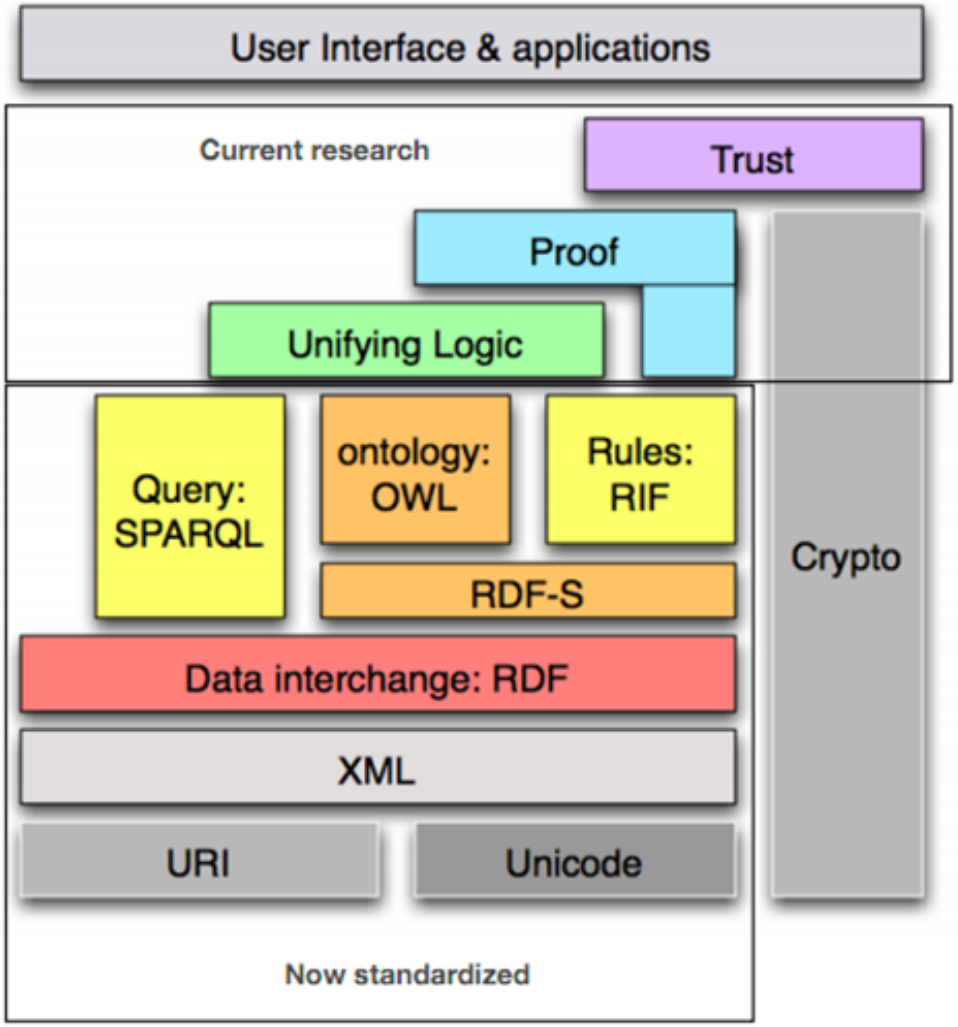
Líneas de trabajo

Soluciones

- a priori: estructurar la información en la Web para facilitar el análisis automático → Web Semántica

- usar métodos de IA, computational statistics, machine learning para analizar la información no estructurada existente en la Web → Knowledge Discovery

Web Semántica



Web Semántica

Se quiere que el significado de la información pueda ser procesada algorítmicamente

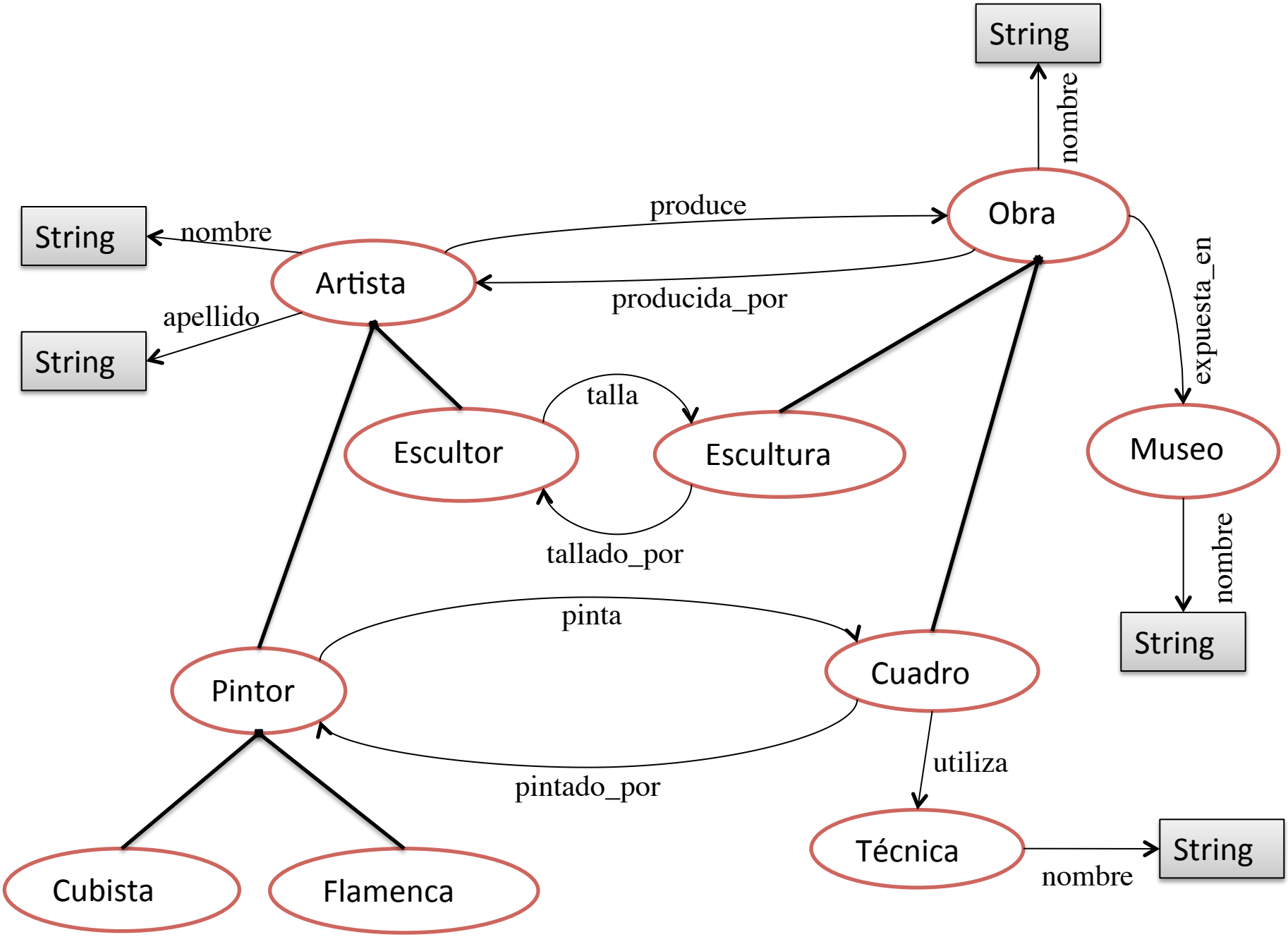
Una forma de lograr lo anterior es mediante la "representación del conocimiento"

- Lógica: proporciona la estructura formal para formular reglas, permitiendo que los algoritmos puedan obtener inferencias
- Ontología: define los objetos, que existen en un dominio particular
- Computabilidad: es una propiedad de una base de conocimientos, que permite que sea efectiva/real

Web Semántica

Ontologías: son representaciones, de un conjunto de conceptos y las relaciones entre ellos en un dominio determinado, lingüísticamente precisas y estructuradas formalmente

Las ontologías se utilizan como medio de estructuración de la información y para el intercambio de datos


String
nombre
Obra
produce
producida_por
expuesta_en
Museo
nombre
String
String
nombre
Artista
apellido
String
talla
Escultor
Escultura
tallado_por
pinta
Pintor
Cuadro
pintado_por
utiliza
Técnica
nombre
String
Cubista
Flamenca

Web Semántica

OWL

- modela muy bien, pero con una inconsistencia todo el razonamiento falla
- al permitir que hayan enlaces pueden aparecer problemas.
- OWL es muy bueno para KR, pero no ha sido "exitoso" para la www
  - más mal uso de sameAs que un buen uso
  - mayor uso de rdf:Class que owl:Class
  - es raro ver que las ontologías se enlacen

Datos enlazados

Usa algunas de las mejores prácticas de la Web Semántica

No se preocupa de tener una ontología "completa"

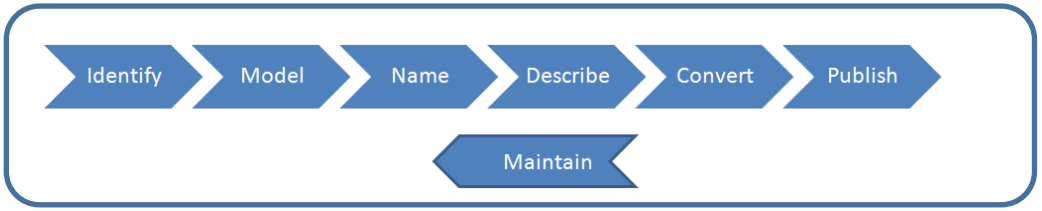
Se enfoca en enlazar

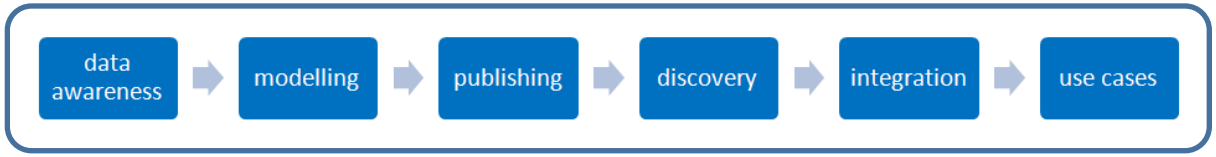
Datos enlazados

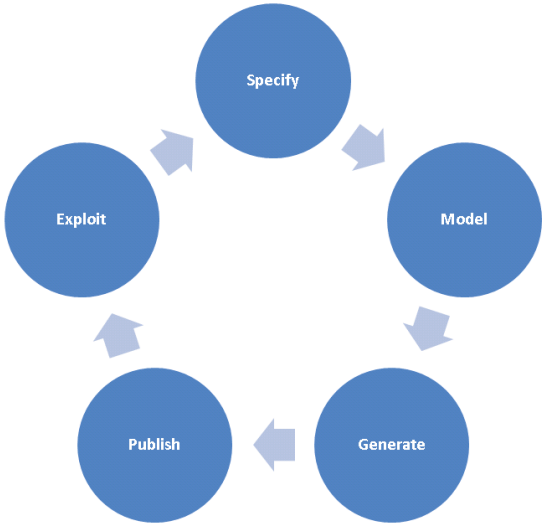
- Use URIs para expresar “cosas”
- Use HTTP URIs para que estas “cosas” puedan ser referenciadas por personas y programas
- Proporcione información útil acerca de la “cosa” (cuando se acceda a la URI) usando estándares como RDF o SPARQL
- Incluya enlaces a otras “cosas” (usando sus URIs)

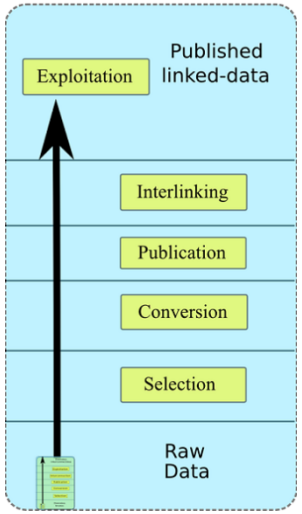
Ciclos de vida, Datos enlazados

http://www.w3.org/2011/gld/wiki/GLD_Life_cycle









Datos enlazados

1. Crear

   extracción de datos, creación de URIs HTTP, seleccionar vocabulario

2. Enlazar

   crear enlaces RDF a datos externos

3. Publicar

   generar los metadatos y dejar disponible el conjunto de datos

Paso 1: crear (extraer los datos)

1. Planillas o datos tabulares

   OpenRefine

2. Bases de datos

   R2RML

3. Textos

   Gate, Stanford NLP, OpenNLP, NLTK, scikit-learn, ANNIE, Wikifier, DBPedia Spotlight, KERT, STOD, PLSA, LDA, etc.

Paso 1: crear (nombrar y diseñar)

1. Todas las cosas o entidades distintas deben tener nombre
2. Diseñar usando Cool Uris http://www.w3.org/TR/cooluris/

Paso 1: crear (buscar vocabularios)

Seleccionar vocabularios para modelar los conceptos y relaciones

- Linked Open Vocabularies
  - http://lov.okfn.org
- Protégé Ontologies
  - http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library#OWL_ontologies
- Open Ontology Repository
  - http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository
- Tones
  - http://owl.cs.manchester.ac.uk/repository/browser
- Watson
  - http://watson.kmi.open.ac.uk/Overview.html
- OBO Foundation Ontologies
  - http://www.obofoundry.org/
- VoCamps
  - http://vocamp.org/wiki/Main_Page
- Falcons
  - http://ws.nju.edu.cn/falcons/objectsearch/index.jsp

Paso 1: crear (buscar vocabularios)

Seleccionar vocabularios para modelar los conceptos y relaciones

- Sindice
  - http://sindice.com/
- SWEO Community Project: Linking Open Data on the Semantic Web
  - http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/CommonVocabularies
- Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets
  - http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/

Paso 1: crear

Obtener un dataset RDF

Resource Description Framework (1998)

Descripción de recursos

- Recurso = identificado por una URI
- Se basa en tripletas
  - Sujeto → Predicado → Objeto

Paso 1: crear

Tripleta RDF

Sujeto
- URI
- Nodo anónimo

Predicado
- URI

Objeto
- URI
- Literal
- Nodo anónimo

Manuel sufre de halitosis

Paso 1: crear

Tripleta RDF

Manuel

Sufre de

halitosis

Paso 1: crear

Tripleta RDF

Manuel	Padece de	halitosis
http://www.example.org/recurso/id/404	http://lexvo.org/id/term/spa/padecer	http://dbpedia.org/resource/Halitosis

Paso 1: crear (extraer los datos)

1. Planillas o datos tabulares

   OpenRefine

2. Bases de datos

   R2RML

3. Textos

   Gate, Stanford NLP, OpenNLP, NLTK, scikit-learn, ANNIE, Wikifier, DBPedia Spotlight, KERT, STOD, PLSA, LDA, etc.

¿Cómo crear datos de Textos?

Data Mining, Text Mining, Information Extraction… ¿Qué se puede extraer de un documento?

Nivel léxico

- Tokenización: extraer tokens de un documento (palabras, separadores, etc.)
- Separar sentencias: conjunto de sentencias para ser procesadas

Nivel lingüístico

- Part-of-Speech: asignar tipos de palabras (sustantivos, verbos, adjectivos, etc.)
- Deep parsing: construir árboles de sintaxis desde sentencias
- Name entity extraction: identificar nombres de personas, lugares, organizaciones, etc.

Nivel semántico

- Resolución de co-referencia: reemplazar pronombres por nombres correspondientes, mezclar diferentes formas de nombres por una sola entidad
- Semantic labeling: asignar identificadores semánticos a nombres considerando desambiguación
- Resumen: asignar importancia a partes de un documento
- Extracción de hechos: extraer hechos relevantes de un documento

¿Cómo crear datos de Textos?

Santiago, dos de diciembre de dos mil catorce.

Vistos:

En autos rol Nº 1.581-2009 del Tercer Juzgado Civil de Talca, la Municipalidad de Talca, representada por su Alcalde don Juan Castro Prieto, deduce demanda en juicio sumario de declaración, determinación y pago de indemnizaciones por servidumbre de apoyo en postes de alumbrado público y de transmisión eléctrica en contra de Compañía General de Electricidad S.A., representada por don Manuel Crisóstomo Solar, a fin que se declare que los postes de alumbrado público y de transmisión de energía eléctrica ubicados en la comuna de Talca, siendo muebles por naturaleza, constituyen inmuebles por adherencia en conformidad con lo previsto en el artículo 568 del Código Civil y, por su actual ubicación, constituyen bienes inmuebles por adherencia de uso público, de acuerdo a lo dispuesto en el artículo 589 del citado Código o, eventualmente, municipales, si así se justificare; que la demandada, no habiendo acreditado en forma legal el dominio sobre los postes e instalaciones de alumbrado público y de transmisión de energía eléctrica que se ubican en la comuna de Talca, carece de todo derecho a ejercer actos de señor y dueño sobre ellos y menos para apropiarse de los ingresos que éstos generan; que el Decreto Supremo N° 197, del Ministerio de Economía, de 14 de octubre de 2004, como, asimismo, los diversos decretos tarifarios anteriores, no pueden ser aplicados en relación con los hechos de esta causa, ya que ellos no han constituido ni constituyen un precepto legal, de manera que su contenido en cuanto dispuso que tanto el convenio de arriendo de apoyo en postes, como su pago, debe hacerse entre los usuarios de telecomunicaciones y las compañías eléctricas concesionarias, todo ello, porque en el caso de autos, los postes se ubican en bienes nacionales de uso público o municipales y éstos han pasado a tener esa calidad por adherencia, de manera que la única facultada por el artículo 5° letra c) de la Ley Orgánica Constitucional para su administración es la Municipalidad de Talca; que la totalidad de los postes emplazados en bienes nacionales de uso público o de propiedad municipal en la comuna de Talca en que se apoyan los claves de energía eléctrica que usa la demandada, como, asimismo, aquellos que están siendo administrados por ella destinando su uso al apoyo de cables de diversas empresas de telecomunicaciones, deben ser administrados por dicha Corporación, estando ésta expresamente facultada, en forma exclusiva, para cobrar las indemnizaciones y arriendos que corresponden al uso que se les está dando; que por lo expuesto la demandada sea condenada a pagar a la demandante las indemnizaciones y rentas ya percibidas por ella y que se determinen conforme el informe pericial que pide se ordene a continuación, incluyendo los intereses legales por la determinación retroactiva. Todo con costas.

La demandada, al evacuar por escrito el traslado conferido en el comparendo de estilo, opuso la excepción dilatoria de corrección del procedimiento y, en cuanto al fondo, solicitó el rechazo de la acción deducida en su contra, controvirtiendo detalladamente cada uno de los fundamentos del libelo, especialmente la propiedad de los postes de distribución de energía eléctrica, circunstancia esta última que impide al Municipio reclamar perjuicios o pagos al respecto. Todo con costas.

El tribunal de primera instancia, en sentencia de veintidós de mayo de dos mil doce, escrita a fojas 1311 y siguientes, rechazó íntegramente la demanda, sin costas.

Se alzaron ambas partes y la Corte de Apelaciones de Talca, en fallo de dieciocho de noviembre de dos mil trece, que se lee a fojas 1468 y siguientes, revocó la sentencia de primer grado sólo en la parte que eximía del pago de las costas a la demandante y, en su lugar, le impuso ese gravamen, confirmando en lo demás la referida sentencia.

En contra de esta última decisión, la demandante deduce recursos de casación en la forma y en el fondo, por haberse incurrido en vicios e infracciones de ley que han influido, en su concepto, sustancialmente en lo dispositivo del fallo, pidiendo que este tribunal la invalide y dicte la de reemplazo que acoja la demanda o bien ordene que el tribunal no inhabilitado que corresponda lo haga en los términos planteados por su parte, con costas.

Se trajeron estos autos en relación para conocer de ambos recursos.

Considerando:

Recurso de casación en la forma:

Primero: Que el recurrente invoca la causal establecida en el artículo 768 N° 5, en relación con el artículo 170 Nros. 2, 3 , 4, 5 y 6, ambos del Código de Procedimiento Civil, es decir, le reprocha al fallo carecer de la enunciación de las peticiones o acciones y sus fundamentos, tanto de la demandante como de la demandada; haber omitido las consideraciones de hecho y

¿Cómo crear datos de Textos?

Liste des termes | Nuage | Statistiques | Structuration | Bigrammes

Candidat de regroupement	Fréquence	Score (Spécificité)	Variantes orthographiques	Matrice
poste	92	358.49	*postes*	Nom
uso público	75	329.3	*uso público*	Nom Adjectif
bien nacional	66	312.72	*bienes nacionales*	Nom Adjectif
eléctrico	142	200.65	*eléctrico* *eléctricos* *eléctricas*	Adjectif
talca	22	177.79	*talca*	Nom
distribución de energía eléctrico	19	164.62	*distribución de energía eléctrica*	Nom Préposition Nom Adjectif
concesión eléctrico	18	159.99	*concesión eléctrica* *concesiones eléctricas*	Nom Adjectif
municipalidad	18	151.77	*municipalidad*	Nom
concesionaria	16	150.3	*concesionaria*	Nom
municipalidades	20	145.52	*municipalidades*	Nom
servidumbres	18	144.69	*servidumbres*	Nom
comuna	15	136.39	*comuna* *comunas*	Nom
concesionarias	14	135.19	*concesionarias*	Nom
postaciones	13	134.47	*postaciones*	Nom
decreto supremo	13	134.47	*decreto supremo*	Nom Nom
energía eléctrico	40	134.17	*energía eléctrica*	Nom Adjectif
distribución de energía	15	125.73	*distribución de energía*	Nom Préposition Nom
general de urbanismo	11	122.79	*general de urbanismo*	Nom Préposition Nom
constitucional de municipalidades	10	116.52	*constitucional de municipalidades*	Nom Préposition Nom
urbanizadores	10	116.52	*urbanizadores*	Nom
bien	121	110.89	*bien* *bienes*	Nom
código civil	9	109.89	*código civil*	Nom Nom
decreto	50	106.44	*decreto*	Nom

¿Cuán parecidos son los documentos?

Un documento se puede representar por miles de atributos, cada uno almacenando la frecuencia de una palabra en particular (vector de frecuencia de términos)

¿Cuán parecidos son los documentos?

Documento	Recurso de casación	Código civil	municipalidad	eléctrico	Decreto supremo	acusado	Error de derecho	Constitución política	juez	Talca
D1	5	0	3	0	2	0	0	2	0	0
D2	3	0	2	0	1	1	0	1	0	1
D3	0	7	0	2	1	0	0	3	0	0
D4	0	1	0	0	1	2	2	0	3	0

¿Cuán parecidos son los documentos?

Similitud coseno es una medida de similitud que se puede usar para comparar documentos, si el valor es más cercano a 1, más parecidos son, un valor de 0 significa que los dos vectores están en 90 grados (ortogonales)

$$\text{sim}(x,y)=x*y/\|x\|\|y\|$$

Usando el ejemplo anterior, x e y son los primeros dos vectores de frecuencia de términos, es decir, x=(5, 0, 3, 0, 2, 0, 0, 2, 0, 0) e y=(3, 0, 2, 0, 1, 1, 0, 1, 0, 1).

$x^t \cdot y = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$

$$\|x\| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$\|y\| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

sim(x, y) = 0.94

De esta manera, esta medida indica que los documentos son bastante similares

Bibliotecas

Kungliga biblioteket

- https://github.com/libris/librisxl/
- http://librisbloggen.kb.se/
- https://twitter.com/librisnytt
- http://devkat.libris.kb.se/ usuario test password test

Library of Congress
British Library
Bibliothèque nationale de France
Deutsche Nationalbibliothek
Biblioteca Nacional de España
BIBFRAME http://www.loc.gov/bibframe/