

Cómo publicar los datos Hacia los datos abiertos enlazados

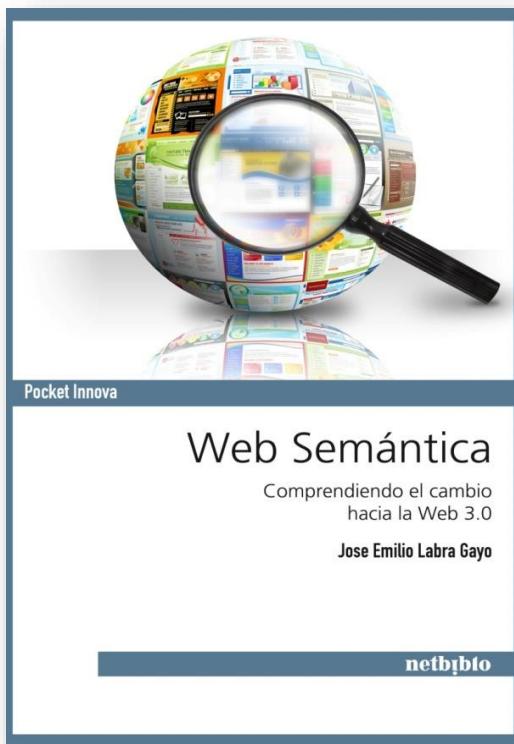
Jose Emilio Labra Gayo
Universidad de Oviedo, España
<http://www.di.uniovi.es/~labra>

Un poco de autobombo

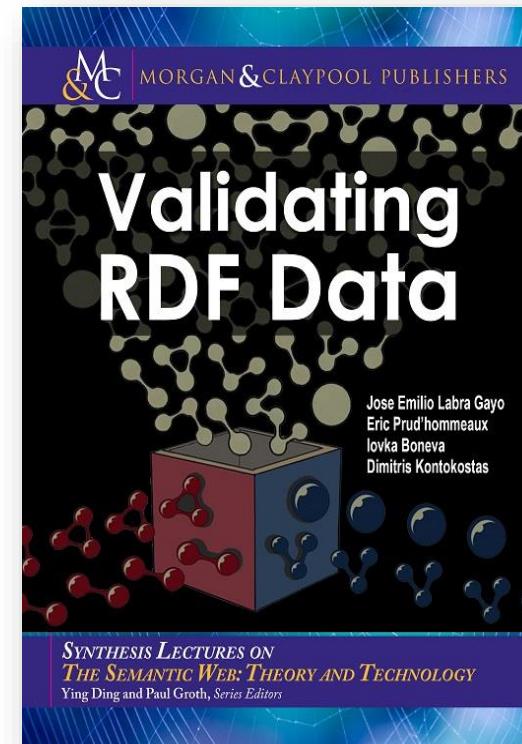
Fundador grupo investigación WESO (WEb Semántica Oviedo)

Autor de libros:

Web Semántica (2012)



Validating RDF Data (2017)



La charla en 1'

¿Qué pasa?

La era de los datos

¿Porqué?

Razones para publicar datos

¿Cómo?

Datos abiertos enlazados





¡Cuidado...llega la
era de los datos!

Avalancha de datos

Producir datos cada vez es más fácil

Tendencias *Open*

Open Software

Open Content

Open Data

Open Science

Open Government

Viejos modelos afectados

Música, Cine, finanzas,...

¿Educación?

¿Gobierno?

...



¿Porqué?

Razones para los gobiernos

Transparencia

Liderazgo

Gobierno como catalizador

Fomentar participación

Nuevas iniciativas y Apps

Razones para los ciudadanos

Nos pertenecen

Creados con dinero público

Queremos mejores servicios



OK, ¡vivan los datos abiertos!
pero...



¿Cómo publicarlos?

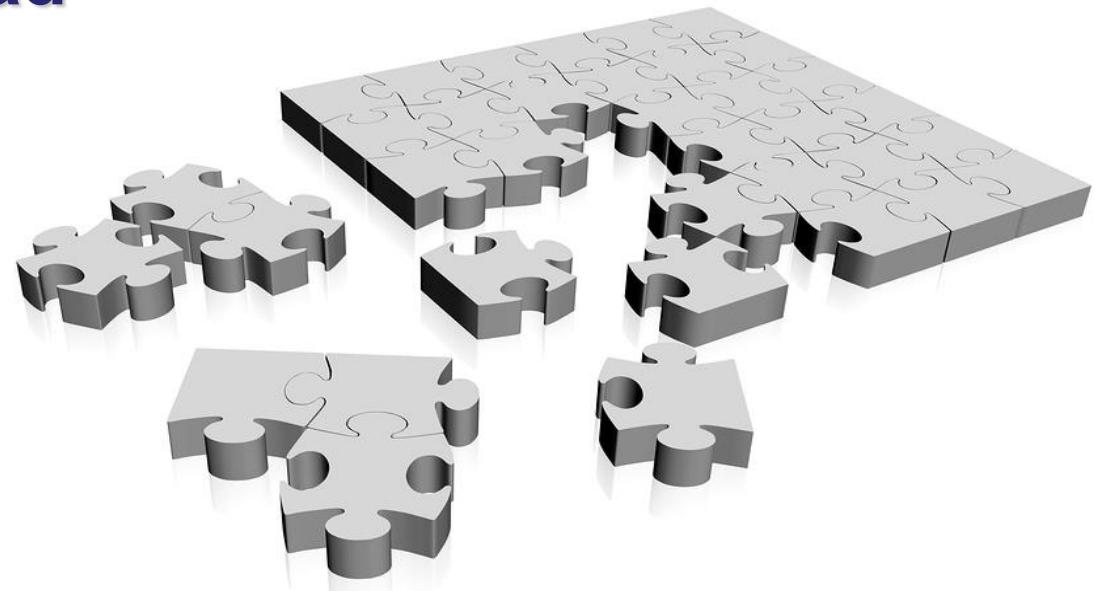
Una reflexión...

Mayor reto informático = Integración

El problema ***no*** es *informatizar* algo

El problema es **integrar** sistemas

Interoperabilidad



Publicar = hacer accesible

Barreras para la accesibilidad

Discapacidad física

Discapacidad técnica (otros entornos)

Intelectual y cultural

Barreras de conocimiento

Otros idiomas

Barreras para las máquinas



¿Accesible para las máquinas?

Sí, es necesario

Accedemos a la web mediante máquinas

Ellas procesan el contenido que vemos

Nos "ayudan" a filtrar, visualizar, etc.



Pero...son entes diferentes

Algunas cosas, fáciles para humanos, difíciles para máquinas

Difícil entender contexto



Ejemplo

"*¿Dónde está Oviedo?*"

Puede ser una ciudad en España

...o una ciudad en Florida, USA

...o un jugador de fútbol

...o....¿Cómo sabemos a qué se refiere?

URLs como identificadores únicos

<http://www.oviedo.es/>

<http://www.cityofoviedo.net/>

https://twitter.com/Bryan_Oviedo

Modelo de Estrellas*

- ★ **Publicar** los datos
(en cualquier formato)
- ★★ Utilizar **formato estructurado**
(Excel en lugar de imágenes escaneadas)
- ★★★ Usar formatos **no propietarios**
(CSV en lugar de Excel)
- ★★★★ Usar **URLs para identificar** datos
(otros sistemas puedan enlazar nuestros datos)
- ★★★★★ **Enlazar con otros** datos externos
(proporcionar contexto)

<http://5stardata.info/>

* Enunciado por Tim Berners-Lee en Gov 2.0 Expo 2010

<http://www.youtube.com/watch?v=ga1aSJXCFe0>



Formatos no estructurados

Formatos binarios o de caja negra

Imágenes, vídeos, música, etc.



Formatos binarios: PDF, PS, etc.



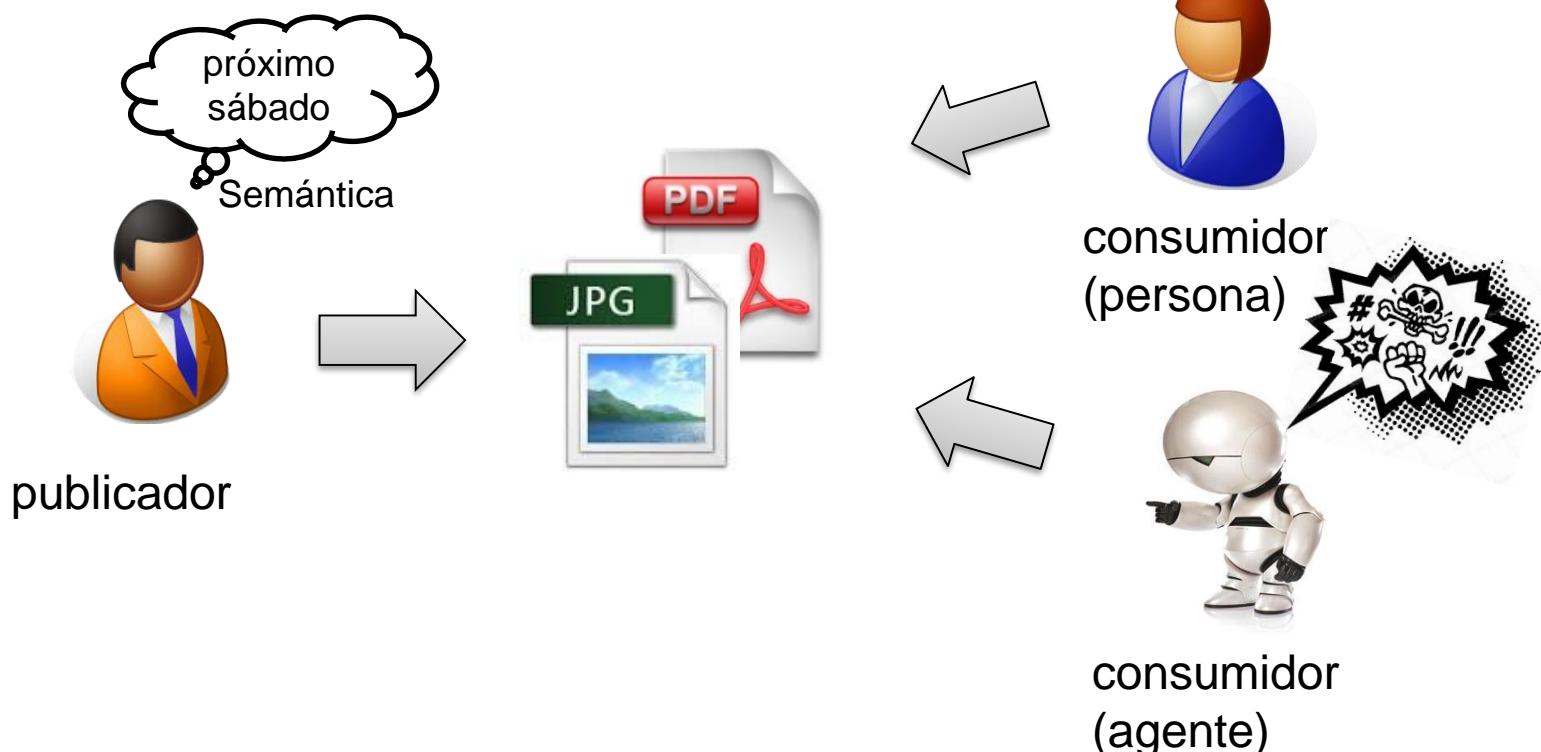
Formatos no estructurados



Problemas

Pérdida de significado

El contenido es difícil de procesar





Formatos estructurados

Los datos tienen una estructura

Ejemplo: Hojas de cálculo

Problema con formatos propietarios

Requieren herramientas que no son públicas





Formatos no propietarios

Utilizar formatos abiertos estructurados

Ejemplos: CSV, HTML

Problema: Contenido depende del contexto



“Comma separated values” valores delimitados por comas

Parados1.csv: Bloc de notas				
Archivo	Edición	Formato	Ver	Ayuda
Municipio;Total;INIC. INDEF.;INIC. TEMPORAL;CONVERT. INDEF.;INIC. INDEF.; ALLANDE;31,;18,;3,;10,;3,;9,;19, ALLER;84,;1,;49,;3,;1,;29,;1,;13,;19,;52, AMIEVA;7,;3,;4,;1,;1,;6, AVILES;1.816,;54,;790,;48,;28,;858,;38,;34,;248,;173,;1.361, BELMONTE DE MIRANDA;25,;3,;22,;1,;1,;23, BIMENES;15,;7,;1,;7,;0,;3,;5,;7, BOAL;6,;2,;1,;3,;6, CABRALES;20,;6,;2,;2,;10,;0,;1,;3,;16, CABRANES;10,;1,;1,;1,;6,;1,;1,;9, CANDAMO;4,;3,;1,;3,;1, CANGAS DE ONIS;88,;1,;30,;3,;3,;47,;4,;1,;3,;14,;70, CANGAS DEL NARCEA;135,;9,;70,;7,;45,;4,;8,;16,;30,;81, CARAVIA;6,;3,;3,;6, CARREÑO;225,;11,;113,;8,;15,;74,;4,;1,;35,;36,;153, CASO;5,;2,;1,;2,;2,;3, CASTRILLON;166,;5,;81,;5,;5,;65,;5,;1,;20,;40,;105, CASTROPOL;48,;36,;1,;2,;8,;1,;2,;26,;7,;13, COANA;27,;1,;11,;4,;0,;10,;1,;1,;6,;4,;16, COLUNGA;59,;23,;2,;33,;1,;2,;1,;11,;45, CORVERA DE ASTURIAS;210,;5,;108,;13,;1,;76,;7,;52,;34,;124, CUDILLERO;66,;13,;32,;2,;2,;16,;1,;14,;9,;13,;30, DEGAÑA;22,;20,;2,;18,;1,;3, FRANCO, EL;30,;16,;14,;3,;1,;10,;16, GIJON;5.824,;113,;2.324,;150,;149,;2.942,;146,;12,;598,;538,;4.676, GOZON;152,;2,;59,;1,;5,;80,;5,;4,;13,;15,;120, GRADO;85,;2,;27,;4,;2,;42,;8,;2,;8,;10,;65, GRANDAS DE SALIME;13,;1,;11,;0,;1,;1,;7,;5, IBIAS;14,;2,;7,;0,;5,;2,;4,;8, ILLANO;2,;2,;0,;1,;1,;1, ILLAS;1,;1,;1,;1,;1, LANGREO;616,;24,;243,;18,;23,;293,;15,;102,;88,;426, LAVIANA;89,;9,;38,;5,;36,;1,;23,;10,;56, LENA;173,;1,;119,;3,;46,;4,;34,;61,;78, LLANERA;620,;21,;297,;36,;20,;230,;16,;1,;73,;68,;478, LLANES;188,;6,;83,;6,;88,;5,;3,;12,;24,;149, MIERES;576,;53,;170,;14,;9,;314,;16,;107,;43,;426, MORCIN;27,;11,;14,;2,;6,;4,;17, MUROS DE NALON;36,;1,;19,;1,;0,;15,;0,;17,;19, NAVA;33,;16,;1,;1,;15,;0,;6,;2,;25, NAVIA;184,;14,;94,;6,;8,;57,;5,;1,;71,;31,;81, NOREÑA;74,;1,;36,;3,;4,;26,;4,;7,;20,;47, ONIS;15,;1,;5,;1,;8,;1,;1,;13, OVIEDO;4.967,;128,;1.591,;95,;159,;2.850,;144,;20,;56,;470,;4.421,				

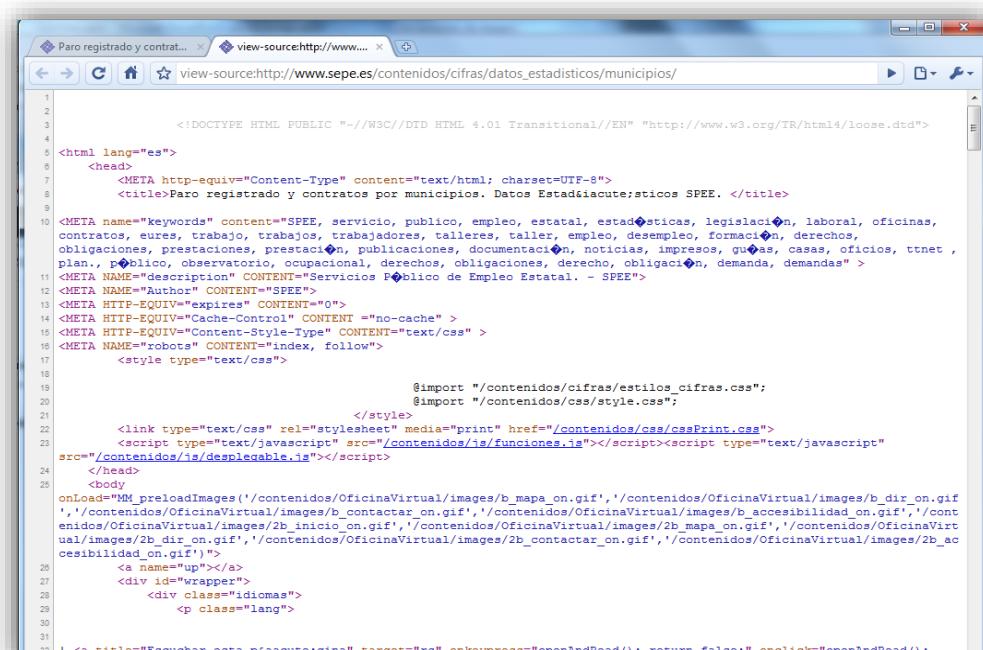


HTML

HTML = representa información que se visualiza en el navegador

Etiquetas h1, p, br, table, ...

Procesar HTML requiere “*screen scrapping*”



The screenshot shows a Microsoft Internet Explorer window displaying the source code of a webpage from the SEPE (Servicio Público de Empleo Estatal) website. The page title is "Paro registrado y contratos por municipios. Datos Estadísticos SPEE". The source code includes standard HTML tags like <html>, <head>, and <body>. It also contains meta tags for keywords, description, and robots, as well as various CSS and JavaScript links. A significant portion of the code is dedicated to managing page layout and accessibility, including the use of 'onLoad' and 'onLoad="MM_preloadImages'" events to load images dynamically.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html lang="es">
  <head>
    <META http-equiv="Content-Type" content="text/html; charset=UTF-8">
    <title>Paro registrado y contratos por municipios. Datos Estadísticos SPEE. </title>
    <META name="keywords" content="SPEE, servicio, público, empleo, estatal, estadísticas, legislación, laboral, oficinas, contratos de trabajo, trabajos, trabajadores, talleres, taller, empleo, desempleo, formación, derechos, obligaciones, prestaciones, prestatción, publicaciones, documentación, noticias, impresos, guías, casas, oficios, ttnet, plan, público, observatorio, ocupacional, derechos, obligaciones, derecho, obligación, demanda, demandas" >
    <META NAME="description" CONTENT="Servicios Públicos de Empleo Estatal. - SPEE">
    <META NAME="Author" CONTENT="SPEE">
    <META HTTP-EQUIV="expires" CONTENT="0">
    <META HTTP-EQUIV="Cache-Control" CONTENT="no-cache" >
    <META HTTP-EQUIV="Content-Type" CONTENT="text/css" >
    <META NAME="robots" CONTENT="index, follow">
    <style type="text/css">
      @import "/contenidos/cifras/estilos_cifras.css";
      @import "/contenidos/css/style.css";
    </style>
    <link type="text/css" rel="stylesheet" media="print" href="/contenidos/css/cssPrint.css">
    <script type="text/javascript" src="/contenidos/is/funciones.js"></script><script type="text/javascript" src="/contenidos/is/desplegable.js"></script>
  </head>
  <body>
onLoad="MM_preloadImages('/contenidos/OficinaVirtual/images/b_mapa_on.gif','/contenidos/OficinaVirtual/images/b_dir_on.gif','/contenidos/OficinaVirtual/images/b_contactar_on.gif','/contenidos/OficinaVirtual/images/b_accesibilidad_on.gif','/contenidos/OficinaVirtual/images/2b_inicio_on.gif','/contenidos/OficinaVirtual/images/2b_mapa_on.gif','/contenidos/OficinaVirtual/images/2b_dir_on.gif','/contenidos/OficinaVirtual/images/2b_contactar_on.gif','/contenidos/OficinaVirtual/images/2b_accesibilidad_on.gif')";
    <a name="up"></a>
    <div id="wrapper">
      <div class="idicmas">
        <p class="lang">
```

URIs para identificar datos



Utilizar URIs para identificar datos

Cada dato tiene una URI diferente

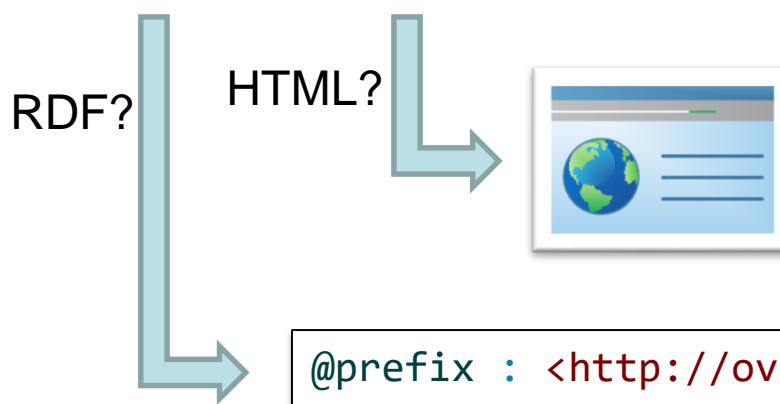
Al acceder a la URI obtenemos representación



Ejemplo: RDF



<http://oviedo.es/monumentos/catedral>



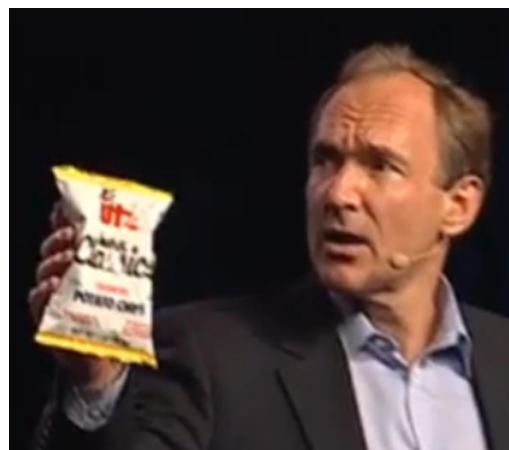
```
@prefix : <http://oviedo.es/monumentos/>

:catedral :municipio "Oviedo" ;
            :arquitecto "Rodrigo Gil de Hontañón" ;
            :fecha      "s. XIII-XVII"
            :estilo     "Gótico" .
```

¿Varias representaciones para lo mismo?



Ejemplo: Bolsa de patatas fritas



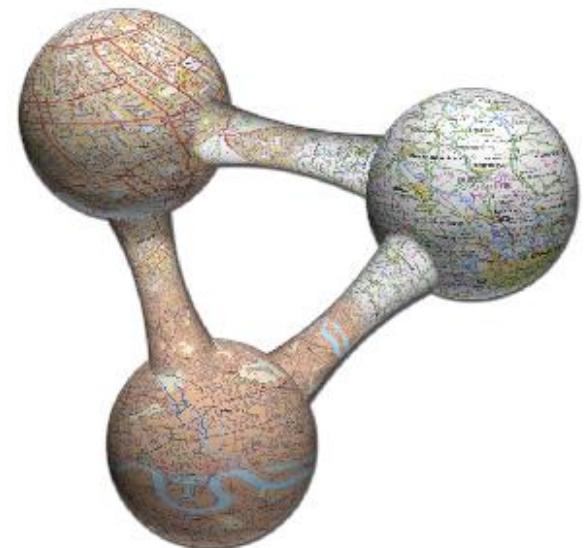
Enlazar con otros datos



Las representaciones devueltas incluyen
enlaces con otros datos

Permite:

Reutilizar y descubrir datos
Aplicaciones "*no previstas*"



Ejemplo: RDF bien enlazado



<http://oviedo.es/monumentos/catedral>

RDF?

HTML?



```
@prefix : <http://oviedo.es/monumentos/>

:catedral dbo:municipality          dbo:Oviedo ;
           dbo:architect            dbr:Rodrigo_Gil_Hontañón ;
           dbo:architecturalStyle    dbr:Gothic_Architecture .
```

```
dbr:Rodrigo_Gil_Hontañón dbo:birthDate "1500-01-01" ;
                           dbo:birthPlace dbr:Rascafria ;
                           dbo:deathPlace dbr:Segovia .
```

Datos abiertos enlazados

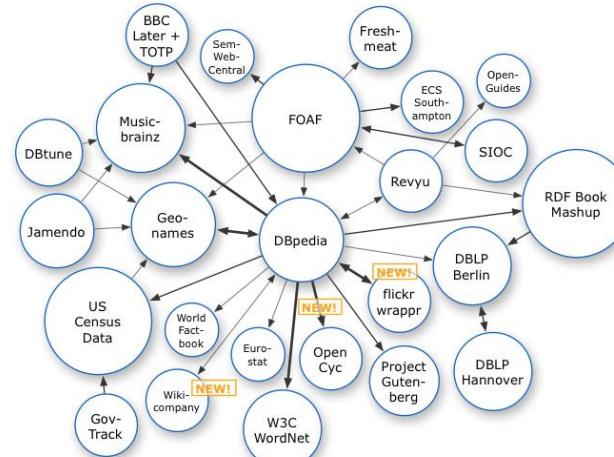
Principios

1. Utilizar URIs para denotar cosas
2. Permitir que las URIs sean dereferenciables
3. Proporcionar información útil
Para personas y máquinas (HTML, RDF)
4. Incluir enlaces a otras cosas relacionadas

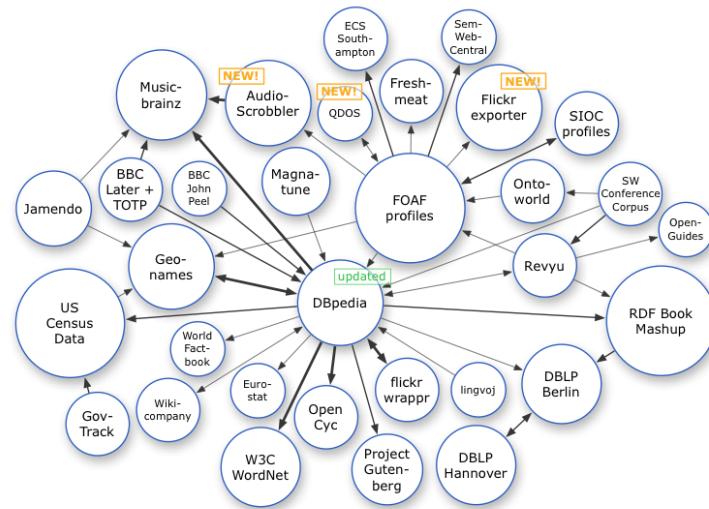
La mejor manera de explotar tus
datos se le ocurrirá a otro

Jo Walsh, Rufus Pollock, http://www.okfn.org/files/talks/xtech_2007/

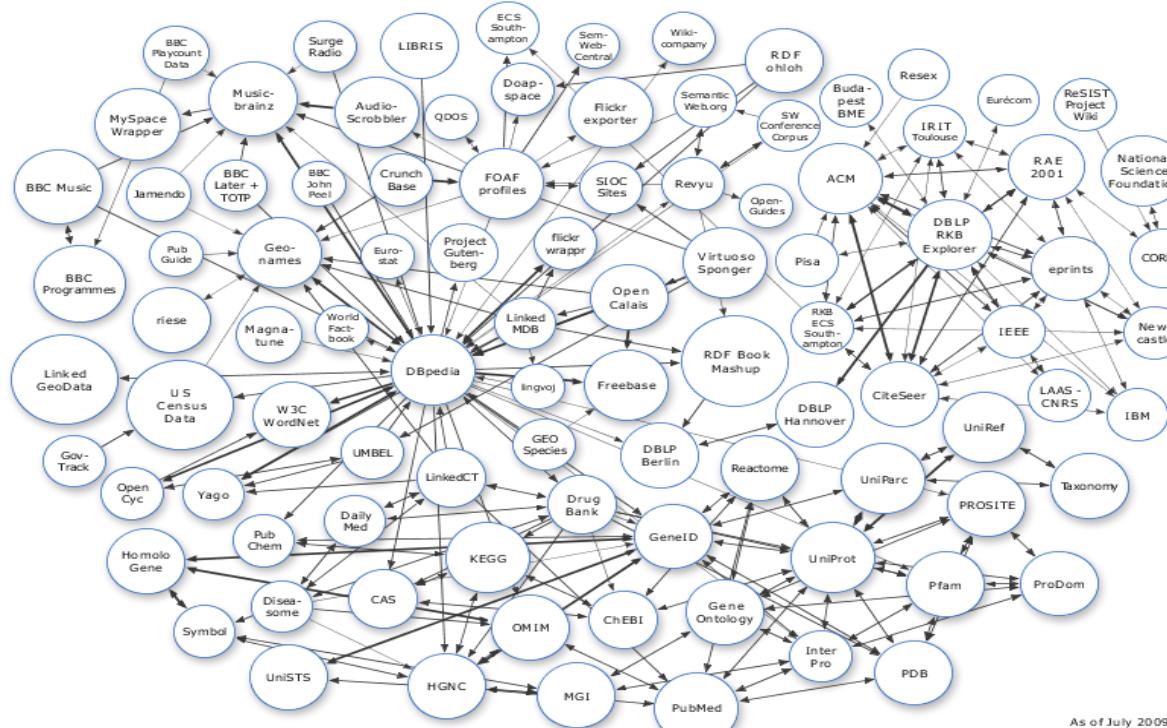
Datos abiertos enlazados (2007)



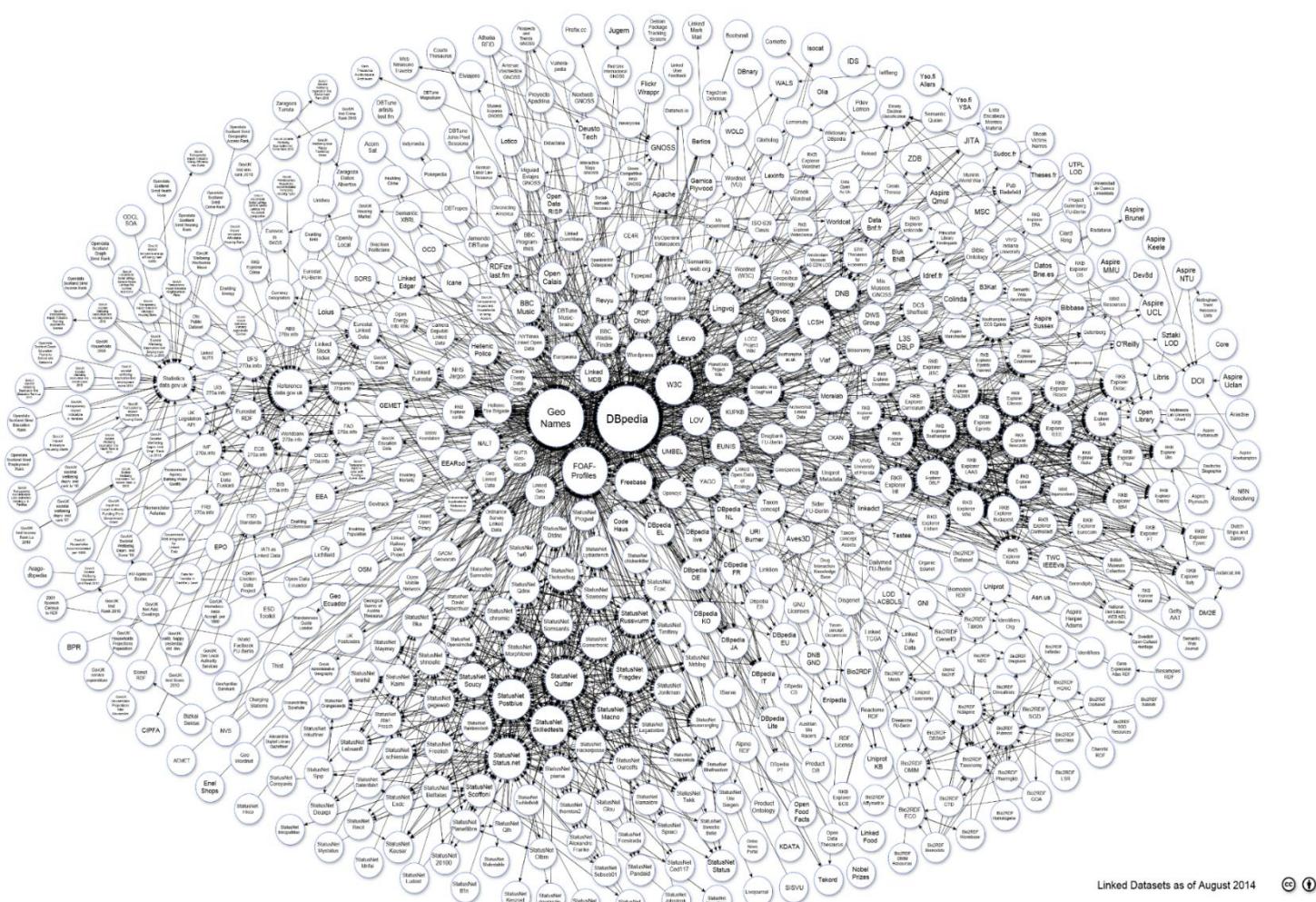
Datos abiertos enlazados (2008)



Datos abiertos enlazados (2009)



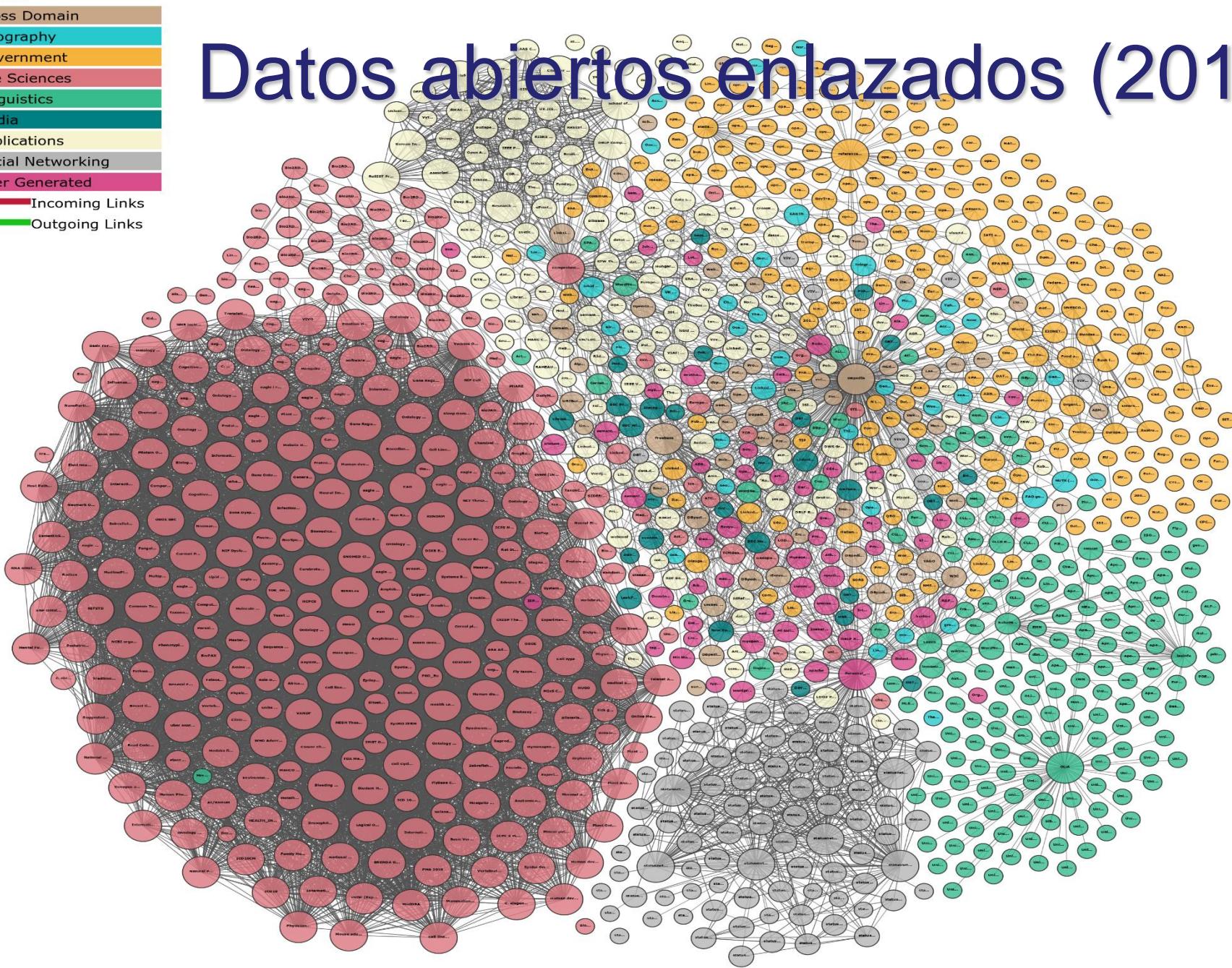
Datos abiertos enlazados (2014)



Linked Datasets as of August 2014



Datos abiertos enlazados (2017)



Fuente: "Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>"

Mitos sobre datos abiertos enlazados

Es muy costoso

Es gratis

Nadie querrá nuestros datos

Ya no está de moda

Demasiada apertura

Va a ser un éxito seguro

Es muy costoso

No necesariamente

Se puede aprender de errores anteriores

Seguir el viejo lema de la web:

Separa contenido de presentación

Contenido: Información/datos

Presentación: aspectos visuales o estéticos

Intentar mantener la semántica



Contenido



Presentación

Es gratis

Pues tampoco

Requiere complementar con visualizaciones

Sólo datos = excesivamente sobrio

Definir modelos de datos y URLs

URLs estables

Contemplar actualización constante

Cuidar las cañerías de datos



Nadie querrá nuestros datos

Al revés...nuestros datos = son nuestro tesoro

Buscadores indexan contenido semántico

Proyecto schema.org (Google, Bing, Yandex,...)

Si facilitemos su trabajo ⇒ mayor posicionamiento

Datos procesables automáticamente = mucho valor

Fomentar cultura de datos

Pueden surgir nuevos negocios y aplicaciones

Gobierno = catalizador: Hackathones y similares...



Ya no está de moda...

Cuidado con las modas en informática

Muchas tecnologías aparecen/desaparecen

Curva de Gartner (2015)



Demasiada apertura

Si de verdad creemos en la transparencia...

Entonces lucharemos por datos reutilizables

Aún así...

Distinguir:

Datos abiertos

Datos enlazados

Datos públicos, datos privados

Datos agregados

Datos parcialmente abiertos

Datos enlazados y cerrados

...



Va a ser un éxito seguro

¡Pues no! Más fracasos que éxitos

Tecnologías inmaduras

Todavía estamos creando tecnologías facilitadoras

Necesario aprender de errores

Ejemplos de problemas

Proyectos a corto plazo

Personas que bloquean los proyectos

Datos no actualizados

Datos poco útiles y no utilizados



En cualquier caso...



**Lo único imposible
es aquello que no intentas**

Fin de la presentación



Más información:
<http://www.di.uniovi.es/~labra>