

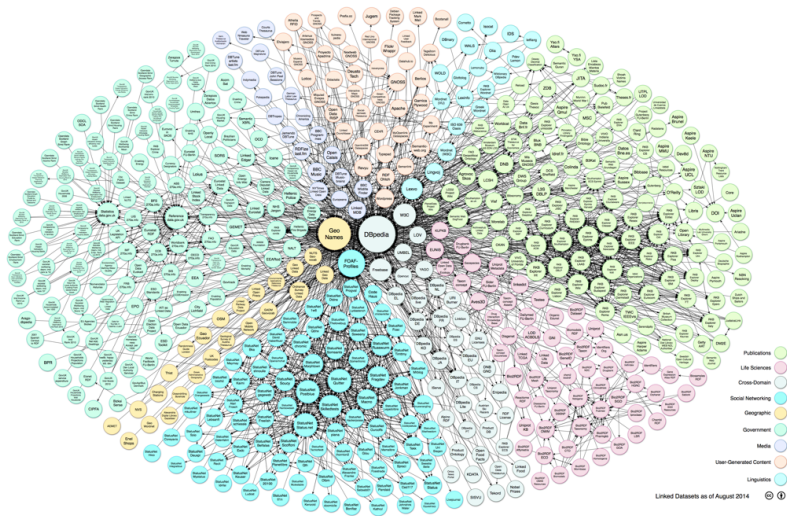
The Semantic Web: A Lightweight Data Integration Platform

Marcelo Arenas
PUC Chile & Center for Semantic Web Research

From the Web of documents to the Web of data



From the Web of documents to the Web of data



Different scenarios: transparency in government data



The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and [more](#).

GET STARTED

SEARCH OVER 192,269 DATASETS



BROWSE TOPICS



Agriculture



Business



Climate



Consumer



Ecosystems



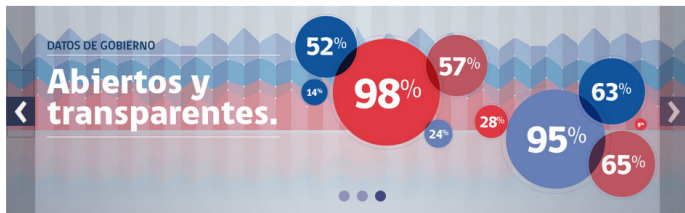
Education



Energy



Different scenarios: transparency in government data



Search bar titled 'Buscar datos' with the placeholder text 'Ej: ambiente'. Below the search bar are several popular tags:

- Etiquetas populares: estadísticas, ejecución presu... Dataset, Salud, presupuesto, educación, gastos, Ley
- ingresos, turismo, seguridad, zonificación, área urbana, PRC, trabajo, previsión, educación superior
- Gobernaciones

Different scenarios: transparency and open data



Visualizaciones



Distribución de compras públicas 2016 por región

Órdenes de compra emitidas durante el 2016, separadas por región en su primer nivel y luego por los organismos

públicos que han transado en cada región como segundo nivel.



Distribución de compras públicas por organismos públicos y proveedores

Órdenes de compra emitidas por cada organismo público y sus respectivos

proveedores adjudicados; separados por año, meses y rango de montos.



Distribución de compras públicas 2015 por sector

Órdenes de compra emitidas durante el 2015, separadas en 7 zonas públicas en su primer nivel y luego por los

organismos públicos que han transado en cada zona como segundo nivel.



Distribución de compras públicas por Convenio Marco

Órdenes de compra emitidas en cada Convenio Marco, separado por los proveedores que han transado en su

primer nivel y luego los organismos públicos que han generado compras a cada proveedor en segundo nivel.

Busca los datos abiertos de las compras públicas

Buscar

Different scenarios: open data

Biblioteca del Congreso Nacional de Chile   **datos.bcn.cl**
Datos Abiertos Enlazados | Portal BCN | Contacto

[Inicio](#) | [Documentación](#) | [Ontologías](#) | [EndPoint SPARQL](#) | [BCN en Linked Open Data](#) | [Acerca de este proyecto](#)

[Inicio](#) > Documentos de BCN open data

Documentos de BCN open data



Datos científicos abiertos: la ciencia la hacemos entre todos.
Manual que explica la política de datos abiertos de CONICYT y entrega recomendaciones para el acceso y preservación de información científica y datos de investigación....
2014
[+ Descargar](#)



Hacia una política integral de gestión de la información pública: todo lo que siempre quisimos saber sobre archivos (y nunca nos animamos a preguntarle al acceso a la información).
Torres, Natalia
2014
[+ Descargar](#)



Ontología de parlamentarios chilenos autores: modelamiento y aplicación.
Quiroz Ubierna, Angela
Seminario para optar al Título de Bibliotecario Documentalista. Universidad Tecnológica Metropolitana....
2013
[+ Descargar](#)



Mándanos tus ideas para enlazar nuestros datos abiertos



Tutoriales recomendados



Las cinco estrellas de los datos abiertos



Datos abiertos en Chile y el mundo

Consuma nuestros datos

[Ver +](#)

Documentos de BCN open data




Datos científicos abiertos: la ciencia la hacemos




Different scenarios: open data


Agency	Article Solution (A)	Maximum Embargo Period	Data Solution (D)
ACL/ NIDILRR	PubMed Central (PMC)	12 months	TBD
AHRQ	PMC	12 months following publication date	Commercial repository, yet to be named* [DMP guidance]
ASPR**	PMC	12 months	Scientific data repositories, data.gov data registry* [DMP guidance]
CDC**	CDC Stacks, using NIHMS submission system	12 months	Multiple solutions + data registry [DMP guidance]
DOD	Defense Technical Information Center (DTIC)	12 months	No specific solution* [DMP guidance]
DOE	Public Access Gateway for Energy and Science (PAGES)	12 months	Varies by office* [DMP template]
DOT	DOT National Transportation Library (NTL)	N/A	To be released [DMP guidance]
FDA**	PMC	12 months	Disciplinary data repositories, where available* [DMP guidance]

Different scenarios: scientific open data





[DATA](#) [TOPICS](#) - [IMPACT](#) [APPLICATIONS](#) [DEVELOPERS](#) [CONTACT](#)

 **SCIENCE & RESEARCH**

[Updates](#) [Apps](#) [Data](#) [Contact Science & Research](#)

Discover high-value public science and research data from across the Federal Government. Need something that you don't see here? [Let us know!](#)


HIGHLIGHTS

Federal R&D Facilities for Entrepreneurs and Innovators

[View this Dataset](#)

As part of the Administration's [Lab-to-Market Initiative](#), agencies are publishing machine-readable data on over 700 Federal R&D facilities that may be utilized by entrepreneurs and innovators to research, prototype, and test new technologies. These facilities, operated by [NASA](#), the [Department of Energy \(DOE\)](#), and the [National Institute of Health \(NIH\)](#), include cutting-edge research tools and together represent billions of dollars of taxpayer investment.

Each facility has its own set of use policies, so a contact person is included in the data wherever possible. For example, some entrepreneurs may be able to access NASA's National Center for Advanced Manufacturing to produce the high-strength, defect-free joints required for cutting-edge aeronautics, or DOE's Manufacturing Demonstration Facility at Oak Ridge National Laboratory for collaborative projects in additive manufacturing, composites and carbon fiber, and other leading clean energy technologies. [Learn more...](#)



Different scenarios: scientific open data

The screenshot shows the Data.Gov website interface. At the top, there is a search bar with the text "Search Data.Gov" and a magnifying glass icon. Below the search bar is the Data.Gov logo, which includes the American flag and the text "DATA.GOV". To the right of the logo are navigation links: "DATA", "TOPICS", "IMPACT", "APPLICATIONS", "DEVELOPERS", and "CONTACT".

Below the navigation links is a blue header bar with the text "SCIENCE & RESEARCH — DATA CATALOG" on the left, a home icon and "Datasets" link in the center, and "Organizations" and a question mark icon on the right. Below this header is a light blue bar with navigation links: "Updates", "Apps", "Data", and "Contact Science & Research".

The main content area has a search bar with the text "Search datasets..." and a magnifying glass icon. To the right of the search bar is a dropdown menu labeled "Order by:" with the text "Select an option". Below the search bar is the text "Datasets ordered by Popular".

Below the text "Topics:" is a button labeled "Research". Below this is a "Filter by location" section with a "Clear" link and a dropdown menu with the text "Enter location...". Below the dropdown menu is a map of the United States with a zoom in (+) and zoom out (-) button on the left and a map icon on the right. Below the map is the text "Map tiles & Data by OpenStreetMap, under CC BY SA".

The main content area displays "54 datasets found". Below this is a list of datasets. The first dataset is "Farmers Markets Directory and Geographic Data" with 777 recent views. It is described as "Department of Agriculture — Longitude and latitude, state, address, name, and zip code of Farmers Markets in the United States" and has an "Excel" download button. The second dataset is "Food Environment Atlas" with 383 recent views. It is described as "Department of Agriculture — Food environment factors—such as store/restaurant proximity, food prices, food and nutrition assistance programs, and community characteristics—interact to..." and has "HTML", "api", and "Excel" download buttons. Both datasets have a green diagonal banner labeled "Federal" to their right.

At the bottom of the page are navigation icons: a home icon, a magnifying glass icon, and a refresh icon.

Different scenarios: scientific open data



Presentación Participación Política Beneficios Buenas Prácticas Noticias Eventos



La Ciencia la hacemos entre todos

Participa con tu opinión, experiencias y conocimiento en la definición de una **Política de datos abiertos** para las investigaciones científicas.

1 of 5

¿Cómo garantizar el acceso a información y datos científicos?

Chile necesita una política de acceso y preservación de información y datos de investigaciones científicas financiadas con fondos públicos. CONICYT ha elaborado [una propuesta](#), pero necesitamos su opinión para perfeccionarla.

Esperamos su opinión



Different scenarios: scientific open data



About ICSU

What we do

Publications

Strengthening international science for the benefit of society



Home > What we do > Interdisciplinary Bodies > World Data System (WDS)

Browse
by subject

International
Research Collaboration

Science
for Policy

Universality
of Science

Online Community

Join

Login

Why join the Online Community ?

World Data System (WDS)

— tags: [International Research Collaboration](#), [Data and Information](#), [University of Science](#), [Science for Policy](#)

About WDS

WDS Scientific Committee

Programme Office

Past Scientific Committees

The ICSU World Data System (WDS) was created through a decision of the 29th General Assembly of the International Council for Science (ICSU).

Building on the 50-year legacy of the ICSU World Data Centres (WDCs) and the ICSU Federation of Astronomical and Geophysical data-analysis Services, the WDS aims at transitioning from existing stand-alone data centres and data services to a common, globally interoperable, distributed data system that incorporates emerging technologies and new scientific data activities.

The new system builds on the potential offered by advanced interconnections between data management components to foster disciplinary and multidisciplinary applications for the benefit of the international scientific community and other stakeholders.

As of 9 September 2016, WDS has 98 Member organizations, including 64 Regular Members, 10 Network Members, 6 Partner Members and 18 Associate Members, and numerous other applications continue to be reviewed by the WDS Scientific Committee (see: www.icsu-wds.org/community/membership)

Different scenarios: scientific open data

LINKED SCIENCE .ORG

EVENTS ▾ PUBLICATIONS TOOLS ▾ VOCABULARIES DATA ▾ PROJECTS ▾ TUTORIALS ▾ SPATIAL JOBS **ABOUT**

ABOUT

Linked Science is an approach to interconnect scientific assets to enable transparent, reproducible and transdisciplinary research.

LinkedScience.org is a community driven-effort to show what this means in practice.

LinkedScience.org was founded early 2011 and is led by [Tomi Kauppinen](#) affiliated 2010-2012 with the [Institute for Geoinformatics](#) at the [University of Muenster](#) (Germany) and since Autumn 2012 with the [Department of Media Technology](#) at the [Aalto University](#) (Finland), and from April 2014 to September 2014 with the University of Bremen (Germany).

Different scenarios: scientific open data



OSDC LOGIN APPLY RESEARCH CLOUDS PUBLIC DATA PIRE HELP

OSDC OPEN SCIENCE DATA CLOUD

A Petabyte-scale Scientific Community Cloud

The OSDC enables scientific researchers to easily manage, share, and analyze large datasets.

[NEW USERS](#) [RETURNING USERS](#)

What is the OSDC? [>> Watch a Video](#)

OSDC in brief

The Open Science Data Cloud provides the scientific community with resources for storing, sharing, and analyzing terabyte and petabyte-scale scientific datasets. The OSDC is a data science ecosystem in which researchers can house and share their own

Why is there a need?

With datasets growing larger and larger, researchers are finding that the bottleneck to discovery is no longer a lack of data but an inability to manage, analyze, and share their large datasets. Individual researchers can no longer download and analyze the

Navigation icons: back, forward, search, and refresh.

Different scenarios: scientific open data

nature.com > scientific data > about

a natureresearch journal

SCIENTIFIC DATA

SEARCH E-alert SUBMIT LOGIN

About


Scientific Data is a peer-reviewed, open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data. [Read our key principles ►](#)

Scientific Data welcomes submissions from a broad range of research disciplines, including descriptions of big or small datasets, from major consortiums to single research groups. Scientific Data primarily publishes Data Descriptors, a new type of publication that focuses on helping others reuse data, and crediting those who share. [Read our aims & scope ►](#)

All content is hosted on nature.com – the destination of millions of scientists globally every month. Publications are indexed in PubMed, MEDLINE and Google Scholar, and are automatically deposited into PubMed Central.

Editors, Advisory Panel and Editorial Board

Scientific Data is supported by a diverse group of researchers, funders and librarians, who form our Advisory Panel and Editorial Board.



What are the common challenges?

What are the common challenges?

- ▶ Storing

What are the common challenges?

- ▶ Storing
- ▶ Publishing

What are the common challenges?

- ▶ Storing
- ▶ Publishing
 - ▶ Anonymizing

What are the common challenges?

- ▶ Storing
- ▶ Publishing
 - ▶ Anonymizing
- ▶ Linking

What are the common challenges?

- ▶ Storing
- ▶ Publishing
 - ▶ Anonymizing
- ▶ Linking
 - ▶ How can we relate pieces of data in different repositories? How can we make these relationships human and machine understandable?
How can we identify the same element in different repositories?

What are the common challenges?

- ▶ Storing
- ▶ Publishing
 - ▶ Anonymizing
- ▶ Linking
 - ▶ How can we relate pieces of data in different repositories? How can we make these relationships human and machine understandable?
How can we identify the same element in different repositories?
- ▶ Integrating

What are the common challenges?

- ▶ Storing
- ▶ Publishing
 - ▶ Anonymizing
- ▶ Linking
 - ▶ How can we relate pieces of data in different repositories? How can we make these relationships human and machine understandable?
How can we identify the same element in different repositories?
- ▶ Integrating
 - ▶ Can we provide a common view of the data? How can we translate data from one repository to another?

What are the common challenges?

- ▶ Querying

What are the common challenges?

- ▶ Querying
- ▶ Ranking

What are the common challenges?

- ▶ Querying
- ▶ Ranking
- ▶ Updating

What are the common challenges?

- ▶ Querying
- ▶ Ranking
- ▶ Updating
- ▶ Visualizing

What are the common challenges?

- ▶ Querying
- ▶ Ranking
- ▶ Updating
- ▶ Visualizing
- ▶ Dealing with uncertainty

What are the common challenges?

- ▶ Querying
- ▶ Ranking
- ▶ Updating
- ▶ Visualizing
- ▶ Dealing with uncertainty

Is there a common framework where these problems can be studied and (hopefully) solved?

What are the common challenges?

- ▶ Querying
- ▶ Ranking
- ▶ Updating
- ▶ Visualizing
- ▶ Dealing with uncertainty

Is there a common framework where these problems can be studied and (hopefully) solved?

- ▶ The Semantic Web is a common framework to tackle these issues

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.”

[Tim Berners-Lee et al. 2001.]

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.”

[Tim Berners-Lee et al. 2001.]

Specific goals:

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.”

[Tim Berners-Lee et al. 2001.]

Specific goals:

- ▶ Build a resource description language with standard semantics

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.”

[Tim Berners-Lee et al. 2001.]

Specific goals:

- ▶ Build a resource description language with standard semantics
 - ▶ Make semantics machine-processable and understandable

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.”

[Tim Berners-Lee et al. 2001.]

Specific goals:

- ▶ Build a resource description language with standard semantics
 - ▶ Make semantics machine-processable and understandable
- ▶ Incorporate logical infrastructure to reason about resources

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.”

[Tim Berners-Lee et al. 2001.]

Specific goals:

- ▶ Build a resource description language with standard semantics
 - ▶ Make semantics machine-processable and understandable
- ▶ Incorporate logical infrastructure to reason about resources

W3C proposals: [Resource Description Framework \(RDF\)](#) and [SPARQL](#)

The Semantic Web



2016

The 15th International
Semantic Web Conference

October 17-21 Kobe, Japan



Home

Calls ▾

Important Dates

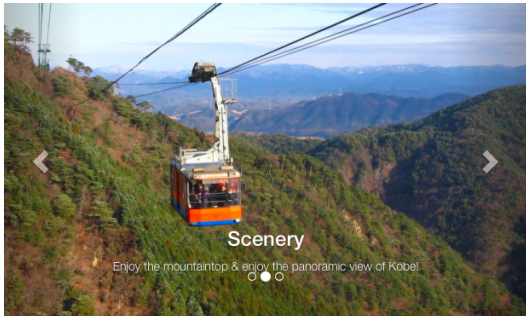
Program ▾

Attending ▾

Organization ▾

Sponsorship

Archives ▾



Scenery

Enjoy the mountaintop & enjoy the panoramic view of Kobe!



@ISWC2016

Important News

- [Awards](#) are announced !
- [AfterFiver](#) is announced !

Sponsors

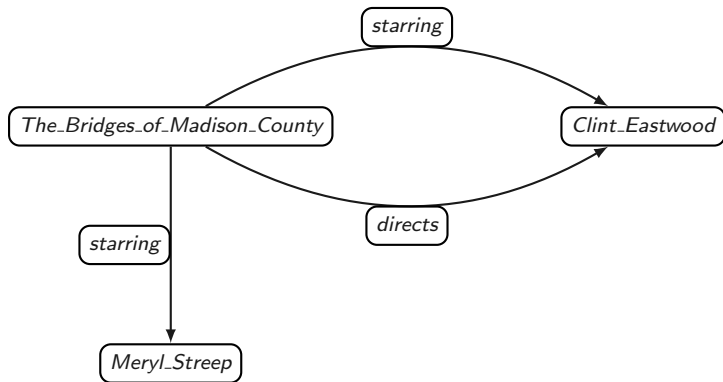
Platinum Sponsors



RDF in a nutshell

- ▶ RDF is the W3C proposal framework for representing information in the Web
- ▶ Abstract syntax based on directed labeled graph
- ▶ Extensible URI-based vocabulary

An RDF graph



An RDF graph in real life: DBpedia

[http://dbpedia.org/resource/The_Bridges_of_Madison_County_\(film\)](http://dbpedia.org/resource/The_Bridges_of_Madison_County_(film))
<http://dbpedia.org/property/director>
http://dbpedia.org/resource/Clint_Eastwood .

[http://dbpedia.org/resource/The_Bridges_of_Madison_County_\(film\)](http://dbpedia.org/resource/The_Bridges_of_Madison_County_(film))
<http://dbpedia.org/property/starring>
http://dbpedia.org/resource/Clint_Eastwood .

[http://dbpedia.org/resource/The_Bridges_of_Madison_County_\(film\)](http://dbpedia.org/resource/The_Bridges_of_Madison_County_(film))
<http://dbpedia.org/property/starring>
http://dbpedia.org/resource/Meryl_Streep .

Prefixes simplify the notation

Prefixes can be defined in an RDF graph to simplify notation

- ▶ They are defined also using triples

Prefixes simplify the notation

Prefixes can be defined in an RDF graph to simplify notation

- ▶ They are defined also using triples

We can include in an RDF graph the following triples:

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

Prefixes simplify the notation

Prefixes can be defined in an RDF graph to simplify notation

- ▶ They are defined also using triples

We can include in an RDF graph the following triples:

```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

Then `<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>` can be replaced by `rdf:type`

How are URIs created and assigned?

How are URIs created and assigned?

There is no centralized mechanism

How are URIs created and assigned?

There is no centralized mechanism

A key component to deal with this issue: `owl:sameAs`

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://cs.dbpedia.org/resource/Meryl_Streepová .
```

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://yago-knowledge.org/resource/Meryl_Streep .
```

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://data.nytimes.com/32250484050106278413 .
```

How are URIs created and assigned?

There is no centralized mechanism

A key component to deal with this issue: `owl:sameAs`

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://cs.dbpedia.org/resource/Meryl_Streepová .
```

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://yago-knowledge.org/resource/Meryl_Streep .
```

```
http://dbpedia.org/resource/Meryl_Streep    owl:sameAs  
http://data.nytimes.com/32250484050106278413 .
```

Reasoning capabilities are needed to deal with `owl:sameAs`

Querying RDF: SPARQL

- ▶ SPARQL is the W3C recommendation query language for RDF (January 2008)
- ▶ Originally it was a graph-matching query language
- ▶ SPARQL 1.1 is the new version of this language, its was released in March 2013

An example of a SPARQL query

Retrieve all the movies in DBpedia

An example of a SPARQL query

Retrieve all the movies in DBpedia

```
?movie rdf:type <http://schema.org/Movie> .
```

An example of a SPARQL query

Retrieve all the movies in DBpedia

```
WHERE
{
  ?movie  rdf:type  <http://schema.org/Movie> .
}
```

An example of a SPARQL query

Retrieve all the movies in DBpedia

```
SELECT ?movie
WHERE
{
    ?movie  rdf:type  <http://schema.org/Movie> .
}
```

Important problems when querying RDF

Important problems when querying RDF

- ▶ Returning as much information as possible

Important problems when querying RDF

- ▶ Returning as much information as possible
- ▶ Reasoning with ontologies

Important problems when querying RDF

- ▶ Returning as much information as possible
- ▶ Reasoning with ontologies
- ▶ Dealing with incomplete information

Important problems when querying RDF

- ▶ Returning as much information as possible
- ▶ Reasoning with ontologies
- ▶ Dealing with incomplete information
- ▶ Exploiting the graph structure of RDF

Important problems when querying RDF

- ▶ Returning as much information as possible
- ▶ Reasoning with ontologies
- ▶ Dealing with incomplete information
- ▶ Exploiting the graph structure of RDF
- ▶ Working with highly distributed data

The Center for Semantic Web Research

An initiative of four universities: PUC Chile, Universidad de Chile, Universidad de Talca y Universidad Técnica Federico Santa María

- ▶ Funded by the Iniciativa Científica Milenio

The Center for Semantic Web Research

An initiative of four universities: PUC Chile, Universidad de Chile, Universidad de Talca y Universidad Técnica Federico Santa María

- ▶ Funded by the Iniciativa Científica Milenio

Members:

- ▶ 13 full-time professors
- ▶ A large number of Ph.D, Master's and undergrad students

The Center for Semantic Web Research

We do research on all the areas mentioned before.

- ▶ With an emphasis on their scientific aspects

The Center for Semantic Web Research

We do research on all the areas mentioned before.

- ▶ With an emphasis on their scientific aspects

But we are also interested in developing prototypes incorporating our results

The Center for Semantic Web Research

We do research on all the areas mentioned before.

- ▶ With an emphasis on their scientific aspects

But we are also interested in developing prototypes incorporating our results

- ▶ Let me show you two of our applied projects . . .

Detecting and measuring the impact of earthquakes using social sensors

Why do we use Social Media?

- ▶ Communicative nature
- ▶ Twitter is largely used in mobile devices
- ▶ Users use Twitter like a news source

What is a social sensor?

- ▶ Each Twitter user is a social sensor
- ▶ Social sensor detects an event and publish a post (without a known pattern)
- ▶ Very noisy type of sensor

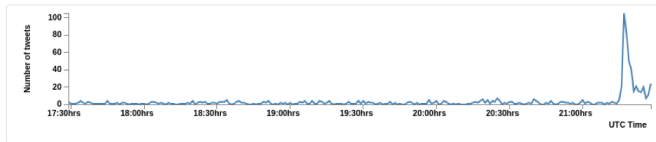
Chile as a study case

- ▶ One of the most seismic countries in the world
- ▶ One of the top-ten highest Twitter users per capita
- ▶ Seismological Center of University of Chile is very interested

What do we want to do?

- ▶ Offer to Seismological Center of University of Chile a tool to explore earthquake related data
- ▶ Complement earthquakes information using tweets data such as location names, opinions about the event and sentiments

Visualization tool



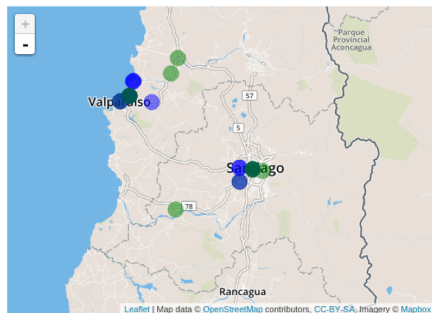
All Short tweets

Carlos Coray 09 May 2016
@RNE_Alpha_8 21:25:00
Sismo 21:16 hrs. Villa Alemana, El Tabo II, La Ligua, Puchuncavi, Rinconada, Valparaíso, Viña del Mar III, Quillota, Zapallar IV Mercalli
Location: Santiago
Reply Retweet Like

Soledad Mora 09 May 2016
@ninatole 21:25:04
Aer que sucede que hay tanto temblor?
Location: Maipú, Santiago
Reply Retweet Like

AlertaNoticiasValpo 09 May 2016
@Alertanoticias

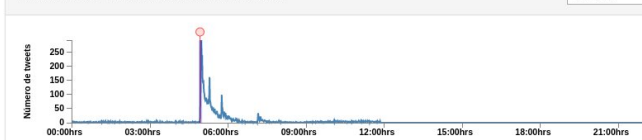
Detailed description: A screenshot of a social media feed showing three tweets. The first tweet is from Carlos Coray (@RNE_Alpha_8) reporting an earthquake (Sismo) at 21:16 hours, listing several locations in Chile and the Mercalli intensity scale. The second tweet is from Soledad Mora (@ninatole) asking 'Aer que sucede que hay tanto temblor?' (What is happening that there is so much tremor?). The third tweet is from AlertaNoticiasValpo (@Alertanoticias). Each tweet includes a location and interaction options (Reply, Retweet, Like).



Visualization tool



Frecuencia de Tweets del 08 Noviembre 2016

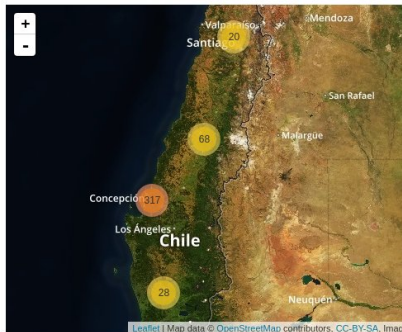


Tweets

Troyano @Dsans404 04:56:05 08/11/16
TERREMOTO
MIERDAAAAAAAAAAAAAAAAAAAA
Localidad del perfil: Gotham city.

Panchito @PanxoGodoy 04:56:05 08/11/16
TERREMOTO
Localidad del perfil: Concepción - Chile

J.L. Escobar @JoseAdición_ 04:56:05 08/11/16
#LoMejorParaCDMX me da asco toda la cdmx ojalá y haya otro terremoto y los mate a todos malditos cambios



Results

- ▶ Currently the system detects earthquakes over 4 in Richter scale
- ▶ It throws alerts in less than 2.5 minutes in the worst case
- ▶ The platform allows to observe Twitter information only a few seconds after an earthquake strikes
- ▶ Geographical visualization has high granularity level although a lot of data are not obtained by GPS

Open data and the Chilean constitutional process

Open data and the Chilean constitutional process

A basis for a new constitution.

- ▶ Effort to create a new Constitution via several small assemblies countrywide

But the data gathered from the process
was not really open

We created a Web repository to:

- ▶ analyze and visualize the (small) data available
- ▶ raise awareness of the importance of opening the data of this process

Web site: <http://constitucionabierta.cl>

Our efforts helped to make the government to open the data of the process

Today there is public access to (anonymized) data of

- ▶ Political opinions of 8.000 groups (of 10-30 people)
- ▶ Individual opinions of 80.0000 participants
- ▶ 2 million “constitutional concepts” prioritized

We have processed this data and cooperated with several public and social organizations

Thank you!

Working with highly distributed data

Web data is highly distributed

Data can be stored in different repositories

- ▶ Different pieces of data have to be collected to answer a query

Web data is highly distributed

Data can be stored in different repositories

- ▶ Different pieces of data have to be collected to answer a query

An important notion to deal with this issue: SPARQL endpoint

- ▶ A Web service that accepts a SPARQL query as input, and returns (part of) the result to the query

Web data is highly distributed

Data can be stored in different repositories

- ▶ Different pieces of data have to be collected to answer a query

An important notion to deal with this issue: SPARQL endpoint

- ▶ A Web service that accepts a SPARQL query as input, and returns (part of) the result to the query

SPARQL has an operator `SERVICE` to query an endpoint

The SPARQL endpoint of DBpedia

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

Execution timeout:

milliseconds *(values less than 1000 are ignored)*

Options:

Strict checking of void variables Log debug info at the end of output (has no effect on some queries and output formats)

(The result can only be sent back to browser, not saved on the server, see [details](#))

Querying DBpedia

We want to retrieve the list of American actors in DBpedia

Querying DBpedia

We want to retrieve the list of American actors in DBpedia

Virtuoso SPARQL Query Editor

[About](#) | [Namespace Prefixes](#) | [Info](#)

Default Data Set Name (Graph IRI)

Query Text

```
SELECT ?name
WHERE
{
  ?actor rdf:type <http://dbpedia.org/class/yago/AmericanFilmActors> .
  ?actor foaf:name ?name .
}
```

The answer to the query

name
"Courtenay Taylor"@en
"Taylor, Courtenay"@en
"Nakia Burrise"@en
"Burrise, Nakia"@en
"Alan Hale, Sr."@en
"Hale, Alan, Sr."@en
"Alec Baldwin"@en
"Baldwin, Alec"@en

...

The SPARQL endpoint of DBLP

We want to retrieve the list of authors in DBLP

The SPARQL endpoint of DBLP

We want to retrieve the list of authors in DBLP

SPARQL:

```
PREFIX d2r: <http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2r-server/config.rdf#>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX map: <file:///home/diederich/d2r-server-0.3.2/dblp-mapping.n3#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
```

```
SELECT ?name
WHERE
{
  ?paper dc:creator ?author .
  ?author foaf:name ?name .
}
```

Results:



The answer to the query

SPARQL results:

name
"Sanjeev Saxena"
"Hans-Ulrich Simon"
"Nathan Goodman"
"Oded Shmueli"
"Norbert Blum"
"Arnold Schönhage"
"Juha Honkala"
"Chua-Huang Huang"
"Christian Lengauer"

...

We would like to combine the previous results ...

```
SELECT ?name
WHERE
{
  ?actor rdf:type <http://dbpedia.org/class/yago/AmericanActors> .
  ?actor foaf:name ?name .
  SERVICE <http://dblp.13s.de/d2r/sparql>
  {
    SELECT ?name
    WHERE
    {
      ?paper dc:creator ?author .
      ?author foaf:name ?name .
    }
  }
}
```

Open issues when dealing with distribution

Some important problems:

- ▶ The notion of SPARQL endpoint needs to be formalized
 - ▶ What queries are accepted?
 - ▶ How is the time distributed between them?
 - ▶ Should a pricing model be used?
 - ▶ What is the protocol to return the answer to a query?

- ▶ A more general notion of endpoint should be formalized and studied

Open issues when dealing with distribution (cont'd)

- ▶ Usability needs to be hugely improved
 - ▶ schema/structure extraction and visualization play a fundamental role here
- ▶ Approaches for discovering relevant data should be studied
- ▶ Operators to distribute the execution of queries should be studied in more depth

Other important problems when querying RDF

Returning as much information as possible

RDF follows an open world assumption

Users may be unaware of the structure of the data

Returning as much information as possible

RDF follows an open world assumption

Users may be unaware of the structure of the data

Thus, the possibility of obtaining additional information if possible is important in this scenario

- ▶ In fact, this feature was present from the very beginning in SPARQL

An optional operator

Retrieve each movie in DBpedia and its gross if this information is available

```
SELECT ?movie ?gross
WHERE
{
  ?movie rdf:type <http://schema.org/Movie> .
  OPTIONAL
  {
    ?movie <http://dbpedia.org/property/gross> ?gross .
  }
}
```

Part of the answer to the query

?movie	?gross
http://dbpedia.org/resource/Frozen_(2013_film)	"1.274E9"
http://dbpedia.org/resource/Amazon_Souls	

What is new?

The `OPTIONAL` operator essentially corresponds to a left-outer join in relational algebra

But ...

What is new?

The OPTIONAL operator essentially corresponds to a left-outer join in relational algebra

But ...

- ▶ The fragments of SPARQL that are natural to study are different than for the case of relational algebra
 - ▶ The complexity of evaluating these fragments was not known [Pérez, A. & Gutierrez 2009; Schmidt, Meier & Lausen 2010]
- ▶ New notions of safeness are needed to avoid a counterintuitive behavior [Pérez, A. & Gutierrez 2009]
- ▶ New optimization techniques are needed [Pérez, A. & Gutierrez 2009; Letelier, Pérez, Pichler & Skritek 2013; Pichler & Skritek 2014]

Reasoning with ontologies

Reasoning capabilities are needed

Reasoning with ontologies

Reasoning capabilities are needed

- ▶ We already mentioned owl:sameAs

Reasoning with ontologies

Reasoning capabilities are needed

- ▶ We already mentioned owl:sameAs
- ▶ An RDF graph can use RDF Schema (RDFS) to establish hierarchies of classes and properties

Reasoning with ontologies

Reasoning capabilities are needed

- ▶ We already mentioned owl:sameAs
- ▶ An RDF graph can use RDF Schema (RDFS) to establish hierarchies of classes and properties
- ▶ The Web Ontology Language (OWL) can be used to define more complex relations between classes and properties

The use of RDFS vocabulary

The following triples are included in DBpedia:

The use of RDFS vocabulary

The following triples are included in DBpedia:

```
http://dbpedia.org/class/yago/Professor110480730
```

```
rdfs:subClassOf
```

```
http://dbpedia.org/class/yago/Academician109759069 .
```

```
http://dbpedia.org/class/yago/Academician109759069
```

```
rdfs:subClassOf
```

```
http://dbpedia.org/class/yago/Educator110045713 .
```

```
http://dbpedia.org/ontology/championInDoubleFemale
```

```
rdfs:subPropertyOf
```

```
http://dbpedia.org/ontology/championInDouble .
```

```
http://dbpedia.org/ontology/championInDouble
```

```
rdfs:subPropertyOf
```

```
http://dbpedia.org/ontology/champion .
```


The use of RDFS vocabulary

Some numbers in DBpedia:

- ▶ triples with `rdfs:subClassOf` as predicate are at least 450K
- ▶ triples with `rdfs:subPropertyOf` as predicate are at least 1K

The use of RDFS vocabulary

Some numbers in DBpedia:

- ▶ triples with `rdfs:subClassOf` as predicate are at least 450K
- ▶ triples with `rdfs:subPropertyOf` as predicate are at least 1K

We need reasoning capabilities to deal with:

- ▶ `rdfs:subClassOf`, `rdfs:subPropertyOf`

The use of RDFS vocabulary

Some numbers in DBpedia:

- ▶ triples with `rdfs:subClassOf` as predicate are at least 450K
- ▶ triples with `rdfs:subPropertyOf` as predicate are at least 1K

We need reasoning capabilities to deal with:

- ▶ `rdfs:subClassOf`, `rdfs:subPropertyOf`
- ▶ and other elements of RDFS such as `rdfs:domain` and `rdfs:range`

Answering a query with RDFS vocabulary

Retrieve all the educators in DBpedia

```
SELECT ?educator
WHERE
{
  ?educator rdf:type
  <http://dbpedia.org/class/yago/Educator110045713> .
}
```

Answering a query with RDFS vocabulary

The answer to the previous query should be the same as for the following query:

```
SELECT ?educator
WHERE
{
  { ?educator rdf:type
    <http://dbpedia.org/class/yago/Academician109759069> . }
  UNION
  { ?educator rdf:type
    <http://dbpedia.org/class/yago/Educator110045713> . }
}
```

Open issues about reasoning with ontologies

Two important problems:

- ▶ Development of efficient query answering algorithms over large RDF graphs with RDFS vocabulary
- ▶ Identification of fragments of OWL that have good expressive power and can be efficiently evaluated

Exploiting the graph structure of RDF

The structure of an RDF graph stores information

It is important to have operators that can deal with this structure

- ▶ In particular, navigating an RDF graph is an important functionality

Exploiting the graph structure of RDF

The structure of an RDF graph stores information

It is important to have operators that can deal with this structure

- ▶ In particular, navigating an RDF graph is an important functionality

Properties paths in SPARQL allow to express reachability queries

Navigating RDF graphs

Get starring actors in the same movie:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?movie <http://dbpedia.org/property/starring> ?actor1 .
  ?movie <http://dbpedia.org/property/starring> ?actor2 .
}
```

Navigating RDF graphs

Previous query can be rewritten by using navigation patterns:

Navigating RDF graphs

Previous query can be rewritten by using navigation patterns:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?actor1 ^<http://dbpedia.org/property/starring> ?movie .
  ?movie <http://dbpedia.org/property/starring> ?actor2 .
}
```

Navigating RDF graphs

Previous query can be rewritten by using navigation patterns:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?actor1
  ^<http://dbpedia.org/property/starring>/
  <http://dbpedia.org/property/starring>
  ?actor2 .
}
```

Navigating RDF graphs

Previous query can be rewritten by using navigation patterns:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?actor1
  ^<http://dbpedia.org/property/starring>/
  <http://dbpedia.org/property/starring>
  ?actor2 .
}
```

The expression in red is called a property path

Navigating RDF graphs

Get starring actors that are connected:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?actor1
  ( ^<http://dbpedia.org/property/starring>/
    <http://dbpedia.org/property/starring> )+
  ?actor2 .
}
```

Navigating RDF graphs

Get starring actors that are connected:

```
SELECT ?actor1 ?actor2
WHERE
{
  ?actor1
  ( ^<http://dbpedia.org/property/starring>/
    <http://dbpedia.org/property/starring> )+
  ?actor2 .
}
```

Can this query be answered?

- ▶ Can it be answered starting from a specific node?

Open issues in exploiting the graph structure of RDF

Some important problems:

- ▶ Development of efficient evaluation algorithms for reachability queries over large RDF graphs
- ▶ Standardization of a query language where paths are first-class citizens