### CC6202-1 LA WEB DE DATOS PRIMAVERA 2015

### **Lecture 9: Linked Data**

Aidan Hogan aidhog@gmail.com

# PREVIOUSLY ...



# (1) Data, (2) Rules/Ontologies, (3) Query

INPUT: "(x, partOf, y)"



But we have not spoken much about ...



... how do we use RDF(S)/OWL/SPARQL to build a "Web of Data"?

### We've gotten from here ...



### To here ...



But what are we missing from here before we can have a "Web of Data"?

## We need links!



... can't have a Web without links!

## We need links!



So how can we include links in this picture?

## RDF filled with IRIs!

@t	base '	<http: <="" th=""><th>//ex:</th><th>1.org/&gt;</th><th><b>&gt;</b> ,</th><th>•</th><th></th><th></th><th></th><th></th><th></th></http:>	//ex:	1.org/>	<b>&gt;</b> ,	•					
<pre>@prefix rdfs: <http: 01="" 2000="" rdf-schema#="" www.w3.org=""></http:></pre>											
<pre>@prefix ex1: <http: #="" ex1.org=""> .</http:></pre>											
<#	Jen>	a <ht< th=""><th>tp://</th><th>/ex1.or</th><th>g,</th><th>/#Perso</th><th>n&gt; ,</th><th>ex1:</th><th>Female</th><th>;</th><th></th></ht<>	tp://	/ex1.or	g,	/#Perso	n> ,	ex1:	Female	;	
	<pre>rdfs:label "Jen"@en ;</pre>		<pre>&lt;#allergy&gt; &lt;#Citrus&gt; ;</pre>								
	ex1:	locati	on [	ex1:la	nt	53.3;	ex1	:long	-9.0	.	

... any IRI could be a link!

## We already saw this in Lab 1 ...

@base <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl> .

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<#me> a foaf:Person;
        foaf:name "Camila Andrea Alvarez"@es;
        foaf:firstName "Camila"@es;
        foaf:surname "Alvarez"@es;
        foaf:birthday "02-24";
        foaf:gender "female";
        foaf:homepage <http://anakena.dcc.uchile.cl/~calvarez/cv.html>;
        foaf:based_near <http://sws.geonames.org/3871336/>;
        foaf:interest <http://dbpedia.org/resource/Tales (series)>,
<http://dbpedia.org/resource/Embroidery>;
        foaf:img <http://images.evisos.cl/2009/06/03/erizo-de-tierra 9de6128c 3.jpg>;
        foaf:schoolHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>;
        foaf:knows <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>,
<http://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
<http://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli> .
```

### These link to other RDF documents ...

@base <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl> .

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<#me> a foaf:Person;
        foaf:name "Camila Andrea Ãlvarez"@es;
        foaf:firstName "Camila"@es;
        foaf:surname "Ãlvarez"@es;
        foaf:birthday "02-24";
        foaf:gender "female";
        foaf:homepage <http://anakena.dcc.uchile.cl/~calvarez/cv.html>;
        foaf:based near shttp://sws.geonames.org/3871336/>;
        foaf:interest <http://dbpedia.org/resource/Tales_(series)>,
 http://dbpedia.org/resource/Embroidery>;
        foaf:img <http://images.evisos.cl/2009/06/03/erizo-de-tierra 9de6128c 3.jpg>;
        foaf:schoolHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>;
        foaf:knows <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>,
<http://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
<http://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli> .
```

ame "Alvarez"@es; hday "02-24"; ler "female"; page <http://anakena.dcc.uchile.cl/~calvarez/cv.html>; d\_near <http://sws.geonames.org/3871336/>; rest <http://dbpedia.org/resource/Tales\_(series)>, rg/resource/Embroidery>; <http://images.evisos.cl/2009/06/03/erizo-de-tierra\_9de6128c\_3.jpg>; olHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>; s <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>, cc.uchile.cl/~jogarrid/web/foaf.ttl#yo>, cc.uchile.cl/~ekauffma/foaf.ttl#eli>.



@base <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

```
<foaf:name "Jaime Salas"@es ;
a foaf:Person ;
foaf:homepage <http://anakena.dcc.uchile.cl/~jasalas> ;
foaf:based_near <http://sws.geonames.org/3871336/> ;
foaf:img <http://anakena.dcc.uchile.cl/~jasalas/web/cat_cage.jpg> ;
foaf:knows <http://anakena.dcc.uchile.cl/~aantoine/web/foaf.ttl#yo> ;
foaf:knows <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl#me> ;
foaf:knows <http://anakena.dcc.uchile.cl/~jmoreno/foaf.ttl#myself> ;
foaf:knows <http://users.dcc.uchile.cl/~sferrada/foaf.ttl#me> ;
foaf:interest "Gaming, programming, lots of -ing." , <http://dbpedia.org/resource/Sharknado</pre>
```

foaf:nick "Ritto" ;
foaf:birthday "25/08" .

#### ame "Alvarez"@es; nday "02-24"; er "female"; page <http://anakena.dcc.uchile.cl/~calvarez/cv.html>; d\_near <http://sws.geonames.org/3871336/>; rest <http://dbpedia.org/resource/Tales\_(series)>; rg/resource/Embroidery>; <http://images.evisos.cl/2009/06/03/erizo-de-tierra\_9de6128c\_3.jpg>; plHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>; < <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>, c.uchile.cl/~jogarrid/web/foaf.ttl#yo>, c.uchile.cl/~ekauffma/foaf.ttl#eli> .

HTTP GET

<pre>@prefix dbo: <http: dbpedia.org="" ontology=""></http:></pre>	•
<pre>@prefix ns1: <http: dbpedia.org="" pre="" resource="" ta<=""></http:></pre>	<pre>les_of_Symphonia:&gt; .</pre>
ns1:_Dawn_of_the_New_World dbo:series	<http: dbpedia.org="" resource="" tales_(series)=""> .</http:>
<pre>@prefix dbr: </pre> <pre></pre> <pre></pre> <pre>@prefix dbr:</pre> <pre></pre>	
dbr:Yoshito dbo:wikiPageDisambiguates	<http: dbpedia.org="" resource="" tales_(series)=""> .</http:>
<pre>@prefix foaf: <http: 0.1="" foaf="" xmlns.com=""></http:> .</pre>	
<pre><http: en.wikipedia.org="" tales_(series)="" wiki=""></http:></pre>	foaf:primaryTopic
<http: dbpedia.org="" resource="" tales_(series)=""> .</http:>	
<pre>@prefix ns4: <http: dbpedia.org="" pre="" resource="" ta<=""></http:></pre>	<pre>les_of_the_Heroes:&gt; .</pre>
ns4:_Twin_Brave dbo:series <http: dbpedia<="" td=""><td>.org/resource/Tales_(series)&gt; .</td></http:>	.org/resource/Tales_(series)> .
<pre>@prefix ns5: <http: dbpedia.org="" ke<="" pre="" resource=""></http:></pre>	roro_RPG:> .
ns5:_Kishi_to_Musha_to_Densetsu_no_Kaizoku	dbo:series
<http: dbpedia.org="" resource="" tales_(series)=""> .</http:>	
<pre>@prefix dbp: <http: dbpedia.org="" property=""></http:></pre>	
dbr:Bandai_Namco_Games dbp:products <http: <="" td=""><td>/dbpedia.org/resource/Tales_(series)&gt; ;</td></http:>	/dbpedia.org/resource/Tales_(series)> ;
<pre>dbo:product <http: dbpedia.org="" pre="" res<=""></http:></pre>	ource/Tales_(series)> .
dbr:Tales_of_Symphonia_Chronicles dbo:ser	<pre>ies <http: dbpedia.org="" resource="" tales_(series)=""></http:></pre>
dbr:Tales_of_Hearts  dbo:series <http: <="" td=""><td>/dbpedia.org/resource/Tales_(series)&gt; .</td></http:>	/dbpedia.org/resource/Tales_(series)> .
dbr:Tales_of_Graces dbo:series <http: <="" td=""><td>/dbpedia.org/resource/Tales_(series)&gt; .</td></http:>	/dbpedia.org/resource/Tales_(series)> .
dbr:Tales of Vesneria dbo:series (http://	/dhredia org/resource/Tales (series)> :

## These also link to other RDF documents ...

@base <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl> .

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<#me> a foaf:Person;
        foaf:name 'Camila Andrea Alvarez'@es;
        foaf:firstName "Camila"@es;
        foaf:surname "Ãlvarez"@es;
        foaf:birthday "02-24";
        foaf:gender "female";
        foaf:homepage <http://anakena.dcc.uchile.cl/~calvarez/cv.html>;
        foaf:based near <http://sws.geonames.org/3871336/>;
        foaf:interest <http://dbpedia.org/resource/Tales (series)>,
<http://dbpedia.org/resource/Embroidery>;
        foaf:img <http://images.evisos.cl/2009/06/03/erizo-de-tierra 9de6128c 3.jpg>;
        foaf:schoolHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>;
        foaf:knows <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>,
<http://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
<http://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli> .
```

foaf:based\_near <nttp://sws.geonames.org/3871336/>;
foaf:interest <nttp://dbpedia.org/resource/Tales\_(series)>,
p://dbpedia.org/resource/Embroidery>;
foaf:img <nttp://images.evisos.cl/2009/06/03/erizo-de-tierra\_9de6128c\_3.jpg>;
foaf:schoolHomepage <nttp://www.uchile.cl/>, <http://www.dcc.uchile.cl/>;
foaf:knows <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>,
p://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
p://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli>.

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ns0: <http://www.w3.org/2003/06/sw-vocab-status/ns#> .
@prefix schema: <http://schema.org/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
foaf: a owl:Ontology ;
  dc11:title "Friend of a Friend (FOAF) vocabulary";
  dc11:description "The Friend of a Friend (FOAF) RDF vocabulary, des
foaf:Person
  a rdfs:Class, owl:Class ;
  rdfs:label "Person" ;
  rdfs:comment "A person." ;
  ns0:term status "stable" ;
```

owl:equivalentClass schema:Person, <http://www.w3.org/2000/10/swap/ rdfs:subClassOf foaf:Agent, geo:SpatialThing ; rdfs:isDefinedBy foaf: ; owl:disjointWith foaf:Organization. foaf:Project .

# PRE-LINKED DATA ...

• Lots of dumps of RDF

Name	Size	Date Modified		
[parent directory]				
README	3.2 kB	10/14/15, 2:00:00 PM		
RELEASE.meta4	10.3 kB	10/14/15, 2:00:00 PM		
RELEASE.metalink	11.5 kB	10/14/15, 2:00:00 PM		
citations.rdf.xz	590 MB	10/14/15, 2:00:00 PM		
core.owl	166 kB	10/14/15, 2:00:00 PM		
databases.rdf.xz	14.1 kB	10/14/15, 2:00:00 PM		
diseases.rdf.xz	576 kB	10/14/15, 2:00:00 PM		
enzyme.rdf.xz	690 kB	10/14/15, 2:00:00 PM		
go.owl.xz	2.9 MB	10/14/15, 2:00:00 PM		
journals.rdf.xz	274 kB	10/14/15, 2:00:00 PM		
keywords.rdf.xz	178 kB	10/14/15, 2:00:00 PM		
locations.rdf.xz	61.7 kB	10/14/15, 2:00:00 PM		
pathways.rdf.xz	49.9 kB	10/14/15, 2:00:00 PM		
proteomes.rdf.xz	3.4 MB	10/14/15, 2:00:00 PM		
taxonomy.rdf.xz	30.1 MB	10/14/15, 2:00:00 PM		
tissues.rdf.xz	34.1 kB	10/14/15, 2:00:00 PM		
uniparc.rdf.xz	52.8 GB	10/14/15, 2:00:00 PM		
uniprot.rdf.xz	61.9 GB	10/14/15, 2:00:00 PM		
uniref.rdf.xz	36.3 GB	10/14/15, 2:00:00 PM		
void.rdf	317 kB	10/14/15, 2:00:00 PM		

### Index of /pub/databases/uniprot/current\_release/rdf/

- Lots of dumps of RDF
- Big OWL ontologies (difficult to re-use)

### **DAML Ontology Library**

Summaries

- Ontologies by URI
- Ontologies by Submission Date
- Ontologies by Keyword
- Ontologies by Open Directory Category
- Ontologies by Class
- Ontologies by Property
- Ontologies by Namespace Used
- Ontologies by Funding Source
- Ontologies by Submitting Organization

#### Queries

- <u>Classes by Name</u>
- Properties by Name

#### Statistics

- 282 ontologies
- <u>Number of Classes</u>, Properties, and Instances Defined
- DAML Property (Feature) Use by Ontology

- Lots of dumps of RDF
- Big OWL ontologies (difficult to re-use)
- No reuse of IRIs ... no links ... no Web!



- Lots of dumps of RDF
- Big OWL ontologies (difficult to re-use)
- No reuse of IRIs ... no links ... no Web!
- Some exceptions, like FOAF
  - but still, blank nodes used more often than IRIs!



# LINKED DATA ...

# Linked Data ... 2006

#### http://www.w3.org/DesignIssues/LinkedData.html

Tim Berners-Lee Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$ Status: personal view only. Editing status: imperfect but published. Up to Design Issues



### Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:



(I'll explain the mug later)

# Four Principles of Linked Data

http://www.w3.org/DesignIssues/LinkedData.html

- Use URIS as names for things 1.
- Use HTTP URIS so that people can look up those names. 2.
- When someone looks up a URI, provide useful information, using 3. the standards (RDF\*, SPARQL)
- Include links to other URIs. so that they can discover more things. 4.



### This is Linked Data ...

@base <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl> .

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
<#me> a foaf:Person;
        foaf:name "Camila Andrea Ãlvarez"@es;
        foaf:firstName "Camila"@es;
        foaf:surname "Ãlvarez"@es;
        foaf:birthday "02-24";
        foaf:gender "female";
        foaf:homepage <http://anakena.dcc.uchile.cl/~calvarez/cv.html>;
        foaf:based near shttp://sws.geonames.org/3871336/>;
        foaf:interest <http://dbpedia.org/resource/Tales_(series)>,
 http://dbpedia.org/resource/Embroidery>;
        foaf:img <http://images.evisos.cl/2009/06/03/erizo-de-tierra_9de6128c_3.jpg>;
        foaf:schoolHomepage <http://www.uchile.cl/>, <http://www.dcc.uchile.cl/>;
        foaf:knows <http://anakena.dcc.uchile.cl/~jasalas/web/foaf.ttl#me>,
<http://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
<http://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli> .
```

### And so we have our "Web of Data"!!!!!









### So what's the catch?

# ONE MINOR CATCH ... ... NAMING *THINGS* NOT DOCUMENTS

# HTTP IRIs usually for documents, not pipes



## My FOAF file is not me

http://aidanhogan.com/foaf.rdf

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:toilet="http://purl.org/az/Toilet#"
xmlns:rel="http://www.perceive.net/schemas/20031015/relationship/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#">

```
<foaf:Document rdf:about="http://aidanhogan.com/foaf.rdf">
<dc:title xml:lang="en">Aidan's FOAF profile</dc:title>
<dc:creator rdf:resource ="http://aidanhogan.com/foaf.rdf#me" />
<dc:date xml:datatype="http://www.w3.org/2001/XMLSchema#date">2005-04-07</dc:date>
<foaf:primaryTopic rdf:resource="http://aidanhogan.com/foaf.rdf#me" />
</foaf:Document>
```

```
<foaf:Person rdf:about="http://aidanhogan.com/foaf.rdf#me">
```

```
<owl:sameAs>
<foaf:Person rdf:about="http://www.deri.ie/about/team/member/Aidan_Hogan#me">
<rdfs:seeAlso rdf:resource="http://www.deri.ie/fileadmin/scripts/foaf.php?id=231" />
</foaf:Person>
</owl:sameAs>
```

```
<owl:sameAs rdf:resource="http://sw.deri.org/~aidanh/foaf/foaf.rdf#Aidan_Hogan" />
<owl:sameAs rdf:resource="http://aidanhogan.com/foaf/foaf.rdf#Aidan_Hogan" />
<owl:sameAs rdf:resource="http://data.semanticweb.org/person/aidan-hogan" />
```

How to identify things not documents?

• We could invent a new protocol real://? real://aidanhogan.com/foaf.rdf



How to identify things not documents

- 1. Hash (aka. fragment identifier)
  - -http://aidanhogan.com/foaf.rdf
    - Identifies my FOAF file
  - -http://aidanhogan.com/foaf.rdf#me
    - Identifies me
    - Look it up with HTTP, you get my FOAF file



How to identify things not documents

- 2. Slash (with 303 redirect)
  - -http://dbpedia.org/data/Sharknado.xml
    - Identifies RDF document about Sharknado
  - -http://dbpedia.org/resource/Sharknado
    - Identifies the movie Sharknado
    - Look it up, it redirects with 303 to RDF doc. above



# Hash vs. Slash



Which is better, hash or slash?

Well, hash has half the number of requests!

GET http://dbpedia.org/resource/Sharknado 303 See Other: http://dbpedia.org/data/Sharknado.xml GET http://dbpedia.org/data/Sharknado.xml 200 Okay: http://dbpedia.org/data/Sharknado.xml



Why does slash exist then?

- Good question
- Hash often used to describe multiple resources in one document
- Slash often used when there are too many resources to fit in one document
  - But nothing would stop, e.g.:
    - http://dbpedia.org/data/Sharknado.xml#this

# So ... erm ... why does slash exist then?



Well, server has more flexibility. It can change which document is returned. In the hash scenario, the document is fixed in the resource IRI.


## Content negotiation with hash





Can also choose from different RDF formats; e.g., Turtle, RDFa, etc. (if supported by the server that is!) Can also specify preferences using q values ...

## Content negotiation with slash

ACCEPT: application/rdf+xml

GET http://dbpedia.org/resource/Sharknado

303 See Other: http://dbpedia.org/data/Sharknado.xml

GET http://dbpedia.org/data/Sharknado.xml

200 Okay: http://dbpedia.org/data/Sharknado.xml





GET http://dbpedia.org/resource/Sharknado

303 See Other: http://dbpedia.org/page/Sharknado

GET http://dbpedia.org/page/Sharknado

200 Okay: http://dbpedia.org/page/Sharknado





# LINKING OPEN DATA

# Linked Data ... 2006

#### http://www.w3.org/DesignIssues/LinkedData.html

Tim Berners-Lee Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$ Status: personal view only. Editing status: imperfect but published. Up to Design Issues



#### Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:



(I'll explain the mug later)

#### Open Data, meet Linked Data



## Linked Open Data



# The 5 $\star$ 's of Linked Open Data

- + Publish data under open licence
- \* \* Make the data "machine readable"
  - e.g., a Spreadsheet better than a PDF table

#### ★★★ Use non-proprietary formats

– e.g., a CSV text file better than Excel

#### • $\star \star \star \star \star$ Use URIs to name your stuff (hint: RDF)

- use unambiguous identifiers that can be linked/looked up
- $\star \star \star \star \star$  Provide links to other content
  - so consumers can follow links to find out more





Oct. 2007



Oct. 2007 Nov. 2007



Oct. 2007 Nov. 2007 Feb. 2008



Oct. 2007 Nov. 2007 Feb. 2008 Sep. 2008



Oct. 2007 Nov. 2007 Feb. 2008 Sep. 2008 Mar. 2009



Oct. 2007 Nov. 2007 Feb. 2008 Sep. 2008 Mar. 2009 July 2009









### Growth of the Linked Data Cloud (sort of)



# THERE IS A LOT OF LINKED DATA OUT THERE NOW ...

## DBpedia: An RDF Encyclopaedia





## Geographical Data: Geonames



## Linked Government Data: data.gov



## Linked Government Data: data.gov.uk

	5 20 101 11	Herre						
🕈 Data App	s Ideas I	Forum Wiki	Blogs	Linked D	ata	Resources	About	
Call for		17 1/4		1	Ove	r 6,900 datas	ets to vi	ew
requests	UN	M 110		2.3	Ins	ide Governm	ient Data	1
Instructions for data publishers	SU UR	190 - 174 - 74 288		33	Who the n data	's who in Governmen noney go? Follow th that opens it all up.	nt and where ese links to t	does înd the
Public Data	1.61	1.00 25.00 10				Government spend o lepartment	ver £25,000,	by
Corporation	91,96	\$12.544 \$44,07 U.W			<b>N</b>	Who does what in W	/hitehall - and	l how
Met office data	Instructions fo	or data publishe	ers		R H	Hospitality, gifts and	expenses	
	Calling all data publis	hers - new guide to pu	blishing to dat	a.gov.uk				
				PAUSE	Sha	re this	Ef	🔊 in
Facts, figures, apps and	more	_			Tama			
					rags			

## Linked Government Data: datos.gob.cl

datos.	gob.cl			¿qué estás buscando?	Q
INICIO	CATĂLOGO	VISUALIZACIONES	APPS	PIDE DATOS	ACERCA DEL SITIO
< A	PROXIMAME SANTIAGO, C	ENTE HILE			

CATÁLOGO

DATASETS PUBLICADOS: 1.232

## Life Sciences



## Life Sciences



Copyright | Privacy | Accessibility | Viewers & Players | Freedom of Information Act | USA.gov U.S. National Library of Medicine | U.S. National Institutes of Health | U.S. Department of Health & Human Services

## Life Sciences



## **E-Commerce**

#### GoodRelations



<<Prev Next>>

#### Main Quickstart Cookbook User's Guide Specification Tools Community

#### The most powerful Web vocabulary for e-commerce

A paradigm shift for e-commerce. Since 2008.

Only 5% of all potential visitors of your site will actually see your offers in their original beauty. 95% will never get beyond a reduced preview of your great products and services as provided by a Web search engine.

GoodRelations is the most powerful vocabulary for publishing all of the details of your products and services in a way friendly to search engines, mobile applications, and browser extensions. By adding a bit of extra code to your Web content, you make sure that potential customers realize all the great features and services and the benefits of doing business with you, because their computers can extract and present this information with ease.

#### Video



#### Who uses GoodRelations?

... and 10,000 more

News from Twitter

#### Contact

#### License

Google Yahoo!

BestBuy

sears.com

kmart.com

See here for additional references.

The GoodRelations ontology is creative

#### Acknowledgments

Many organizations and individuals have supported

Univ.-Prof. Dr. Martin Hepp

## New York Times Meta-data



http://data.nytimes.com/schools/schools.html

## **BBC** Music







# ONLY PROBLEM IS FEW PEOPLE ARE USING THE DATA ...

# ENTERING UNKNOWN TERRITORY: OPEN RESEARCH QUESTIONS!

# CATCH NUMBER 1: INTEGRATION

## **Need for Integration**



http://dbpedia.org/resource/Bill\_Clinton

http://rdf.freebase.com/ns/en.bill\_clinton

http://data.nytimes.com/clinton\_bill\_per

http://www.bbc.co.uk/music/artists/...

How could OWL help here?

owl:sameAs



# A (Hypothetical) Integration Example

What is this query asking?

Find the top cited authors based in Chile, only including papers from DBLP, excluding self-citations from the count.

# A (Hypothetical) Integration Example



#### Marcelo Arenas

Professor of Computer Science, PUC Chile Database theory - applications of logic to computer science - semantic Web Verified email at ing.puc.cl Homepage

Ci	tation in	dices	Citations		
	All	Since 2009	644		
Citations	4356	2488			
h-index	30	25			
i10-index	47	44	0 2000 2002 008 2010		2014
			Show: 20	• 1-20	Next >
Title / Autho	or		c	ited by	Year
Consisten M Arenas, I Proceeding	t query Bertoss s of the e	answers in in i, J Chomicki eighteenth ACM	consistent databases SIGMOD-SIGACT-SIGART symposium on	584	1999
A normal M Arenas, I ACM Trans	form for Libkin actions o	XML docum	ents .tems (TODS) 29 (1), 195-232	469	2004



**₽** 



**Universität Trier** 

#### Marcelo Arenas 💩 🐅 👳

List of publications from the <u>DBLP Bibliography Server</u> - <u>FAQ</u> Other views: <u>by type</u> - <u>by year</u> (modern) - <u>classic-C</u>

Ask others: ACM DL/Guide - S - CSB - MetaPress - Google - Bing - Yahoo author:marcelo\_arenas

Facets and more with CompleteSearch

2013	Refine by AUTHOR		
Marcelo Arenas, <u>Pablo Barceló, Ronald Fagin, Leonid Libkin</u> : Solutions and query rewriting in data exchange. <u>Inf. Comput. 228</u> : 28-61 (2013)	<u>Marcelo Arenas</u> (101) <u>Jorge Pérez</u> (25) <u>Leonid Libkin</u> (22) <u>Pablo Barceló</u> (15) [top 4] [top 50] [all 60]		
Marcelo Arenas, <u>Jorge Pérez, Juan L. Reutter</u> : Data exchange beyond complete data. <u>J. ACM 60</u> (4): 28 (2013)	Refine by VENUE <u>PODS</u> (10) <u>CoRR</u> (7) <u>Encyclopedia of Database Systems</u> (6) <u>Description Logics</u> (4) [top 4] [top 50] [all 52]		



#### Co-authors (93)

Leonid Libkin

Leopoldo Bertossi

Claudio Gutierrez (Claudio Gutiérrez)

Pablo Barcelo (Pablo Barceló)

Jan Chomicki



#### Academic > Authors > Marcelo Arenas





Contents Featured content Current events Random article Donate to Wikipedia Wikimedia Shop


A (Hypothetical) Integration Example































Article ⊑ Paper dblp:Article rdfs:subClassOf dblp:Paper .





affiliation o country ⊑ basedIn ex:basedIn owl:propertyChainAxiom( ex:affilition ex:country)



#### Not clear yet how to do this!





• Not clear how to do reasoning on the Web!

### Side Note: Fact or Fiction?





# **BIG CATCH NUMBER 2: DATA ACCESS**

#### Access Methods

- Client has a request/query Q
- Server has a dataset D
- Client issues Q to server
- Server computes and returns response Q(D)



#### Access Methods



- Multiple clients / multiple servers (blurred)
- Remote, decentralised, uncoordinated
- Web scale

### Linked Data Access Methods

#### 1. Dereferencing:

- Look up a URI, get an RDF document
- 2. Dumps:
  - Get all data in an archive
- 3. SPARQL Queries:
  - Send a query, get the answers

### Dereferencing (what is it?)

### Q = "http://dbpedia.org/resource/Columbia" Q(D) =

- \vector{\vetor{\\vetor{\vetor{\vetor{\vetor{\vetor{\vetor{\vetor{\veto
- xmlns:grs="http://www.georss.org/georss/">
- v<rdf:Description rdf:about="http://dbpedia.org/resource/Afro-Latin\_American'
   <dbpedia-owl:populationPlace rdf:resource="http://dbpedia.org/resource/Col
   </rdf:Description>
- v<rdf:Description rdf:about="http://dbpedia.org/resource/Alb%C3%A1n">
   <dbpedia-owl:country rdf:resource="http://dbpedia.org/resource/Colombia"/>
   </rdf:Description>
- v<rdf:Description rdf:about="http://dbpedia.org/resource/Andr%C3%A9s\_Pastrana <dbpedia-owl:birthPlace rdf:resource="http://dbpedia.org/resource/Colombia <dbpedia-owl:country rdf:resource="http://dbpedia.org/resource/Colombia"/> <dbpprop:birthPlace rdf:resource="http://dbpedia.org/resource/Colombia"/> <dbpprop:placeOfBirth rdf:resource="http://dbpedia.org/resource/Colombia"/> </rdf:Description>
- v<rdf:Description rdf:about="http://dbpedia.org/resource/Antanas\_Mockus">
   <dbpedia-owl:birthPlace rdf:resource="http://dbpedia.org/resource/Colombia
   <dbpprop:birthPlace rdf:resource="http://dbpedia.org/resource/Colombia"/>
   </rdf:Description>

### Dereferencing (what's wrong with it?)

#### Responses vary from server to server

- local triples where URI is **subject** (83%) vs.
- local triples where URI is **subject** or **object** (55%)

WELL-DEFINED: For a given	VELL-DEFINED	
Q and D, chefts and	P	
servers agree on what	Dereferencing	
Q(D) should be.		

### Dereferencing (what's wrong with it?)

Very coarse:

wasting bandwidth

- Give me all capitals of South American countries.
  - Dereference documents for all country URIs
  - See which ones are in South America, throw away rest
  - Throw away triples other than capitals



Dereferencing (what's wrong with it?)

- No pagination:
  - Give me some information about Italy.
    - Load document with 100,000 triples
    - Throw away 99,900 triples the user won't read

PAGINATION: A large response Q(D) can be split		WELL-DEFINED	GRANULAR	PAGINATION		
into chunks	Dereferencing					

#### Dumps (what are they?)



DBpedia Blog | Get Involved | Get Help

#### **DBpedia 3.9 Downloads**

About / News Applications Use Cases Datasets Online Access DBpedia Live Downloads Interlinking

This pages provides downloads of the DBpedia datasets. The DE are licensed under the terms of the Creative Commons Attribution and the GNU Free Documentation License. **DEENION** The dow as N-Triples and N-Quads, where the N-Quads version contains : information for each statement. All files are bzip2 <sup>1</sup> packed. In addition to the RDF version of the data, we also provide a table of the core DBpedia data sets as CSV and JSON files. See DBp Older Versions: DBpedia 3.8, DBpedia 3.7, DBpedia 3.6, DBpedia DBpedia 3.4, DBpedia 3.3, DBpedia 3.2, DBpedia 3.1, DBpedia DBpedia 2.0

### Dumps (what's wrong with them?)

- 15× compression for RDF achievable
- But same weaknesses as for deref. still apply



### Dumps (what's wrong with them?)

- 15× compression for RDF achievable
- But same weaknesses as for deref. still apply
- Also, no standard access methods:
  - Various compressions and formats
  - Linked through generic HTML

ACCESSIBLE: The protocol and formats are defined for automatic access by software agents



### SPARQL (what is it?)

...

*Q* =

PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?capital
WHERE {
 ?s a dbo:Country ; dbp:capital ?c ;
 dcterms:subject category:Countries\_in\_South\_America .
 ?c rdfs:label ?capital . FILTER (lang(?capital)="en")

### Q(D) =

capital		
'Caracas"@en		
'Buenos Aires"@en		
'Asunción"@en		
'Brasília"@en		
'Georgetown, Guyana"@en		
'Montevideo"@en		
'Paramaribo"@en		
'Bogotá''@en		
'Lima"@en		
'Quito"@en		
'Santiago"@en		

#### SPARQL (to the rescue?)

ACCESSIBLE Well-defined Granular

PAGINATION

Dereferencing Dumps SPARQL endpoints





- Single Protocol and RDF Query Language
- SPARQL evaluation: PSpace-complete



CACHEABLE: Common requests can be cached and re-used. Queries can be anticipated.





**COSTABLE:** The cost of processing a query can be anticipated before actual processing.

№ of Results	$N_{2}$ of Endpoints
500	1
1,000	3
1,500	1
5,000	1
10,000	49
20,000	2
40,000	1
50,000	3
100,000	7
Total:	68

#### SELECT \* WHERE { ?s ?p ?o } LIMIT 100002

Virtuoso 42000 Error The estimated execution time 0 (sec) exceeds the limit of 3000 (sec).

```
SPARQL query:
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX dbp: <http://dbpedia.org/ontology/>
SELECT ?capital
WHERE {
  ?s a dbo:Country ;
    dbp:capital ?c ;
    dcterms:subject
category:Countries_in_South_America .
  ?c rdfs:label ?capital .
  FILTER (langMatches(?c,"en"))
```



- Protocol always expects a perfect answer
  - No support for partial results, timeouts, exception handling, pagination ...





• *D* is a black-box for the user

*Q* =

PREFIX dbo: <http://dbpedia.org/ontology/> SELECT (COUNT(?c) as ?count) WHERE { ?c a dbo:Country .



TRANSPARENT: The client can determine if a dataset D is relevant and the service sufficient.


**Problem Categories** 

- 1. Standardised
- 2. Bandwidth-efficient
- 3. Server-processing-efficient
- 4. Usable by client

Dereferencing Dumps SPARQL endpoints Data Access: Open Problem

- Data access methods crucial
  - Access = Protocol + Query Language
- But current ones don't work
  - $\rightarrow$  need something else (or multiple things?)
  - Open question:



# **BIG CATCH NUMBER 3: DATA QUALITY**

### Can't trust everything you read on the Web









New York









Relationship between Coeliac and Gluten?

- Coeliac(Jen)
- Coeliac ≡ ∃allergy.{Gluten}
- ∴ allergy(Jen,Gluten)

[Jen is a Coeliac] [a Coeliac has allergy to Gluten] [Jen has allergy to Gluten]





- Coeliac(Jen)
  Coeliac = ∃allergy.{Gluten}
  CoveS
  Allergy(Jen,Gluten)

[Jen is a Qoeliaces beer [a Coeliac has allergy to Gluten] Jen has allergy to Gluten



# ACTUALLY THERE ARE LOTS OF OPEN ISSUES

## **Open Issues / Research Questions**

- How to efficiently access Linked Data?
- How to automatically link datasets?
- How to reason over Web data?
- How to verify/measure quality?
- How to deal with deceit?
- How to make it all "easy to use"?
- How best to model vocabularies for re-use?
- How to convert legacy data to Linked Data?
- •

None of these problems is a deal-breaker. All of these problems are subject to research! Solutions are being proposed!



### Linked Data

- Four principles:
  - Use URIs (IRIs) to name things
  - Use HTTP URIs that can be looked up
  - Return RDF about that thing when looked up
  - Provide links in the RDF to other RDF documents
- Identifying things, not documents:
  - Hash: use a fragment identifier
  - Slash: use a 303 See Other redirect

### Linked Data

...

• Linked Open Data:

– Publishing Open Data as Linked Data

- Many open issues:
  - How to integrate/link data
  - How to apply reasoning
  - How to access data efficiently
  - Dealing with data quality



Structure of the course

- Marking structure:
  - 40% labs
  - 25% project
  - 5% reading group
  - 30% exam

## Speculative Schedule ③ (NOT DEFINITE!)

- Now: Week 10(1)
- 10(2): Lab
- 11(1): Reading Group / Project Preparation (maybe short lecture)
- 11(2): Lab
- 12(1): Reading Group Presentations
- 12(2): Project
- 13(1): Lecture or Project (not sure yet)
- 13(2): Project
- 14(1): Lecture or Project (not sure yet)
- 14(2): Project Presentations
- 15(1): Lecture (Recap/Revision)
- 15(2): Informal lab (no marking week 15)

