CC6202-1
La Web de Datos
Primavera 2016

Lecture 9: Linked Data

Aidan Hogan
aidhog@gmail.com

PREVIOUSLY …

First SPARQL (1.0)
Then SPARQL 1.1

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

INPUT: "$(x$,partOf,$y)$"

DATA:

`http://ex.org/Ireland`

Ireland

(Ireland,partOf,Europe)
(Ireland,a,Country)
(Ireland,capital,Dublin)

`http://ex.org/Dublin`

Dublin

(Dublin,population,1000000)

RULES: $(a, \text{capital}, b) \rightarrow (b, \text{partOf}, a)$
$(c, \text{partOf}, d), (d, \text{partOf}, e) \rightarrow (c, \text{partOf}, e)$

OUTPUT: $\{(x \mapsto \text{Ireland}, y \mapsto \text{Europe}), (x \mapsto \text{Dublin}, y \mapsto \text{Ireland})$
$(x \mapsto \text{Dublin}, y \mapsto \text{Europe})\}$

But we have not spoken much about …

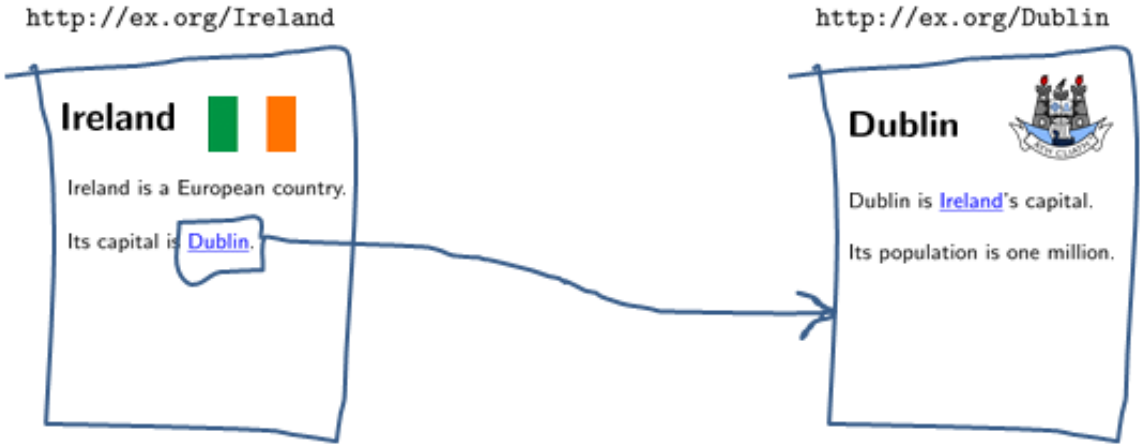
CC6202-1

LA WEB DE DATOS

… how do we use RDF(S)/OWL/SPARQL to build a “Web of Data”?

We've gotten from here …

The current Web is document-centric




```
<html>
 <body>
  <h1>Ireland</h1>
  <div class="flag"><img src="flag.jpg" /></div>
  <p>Ireland is a European country.</p>
  <p>Its capital is <a href="http://ex.org/Dublin">Dublin</a>.</p>
  ...
 </body>
</html>
```

To here …

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

INPUT: "$(x$,partOf,$y)$"

DATA:

`http://ex.org/Ireland`

Ireland 

(Ireland,partOf,Europe)
(Ireland,a,Country)
(Ireland,capital,Dublin)

`http://ex.org/Dublin`

Dublin 

(Dublin,population,1000000)

RULES: $(a, \text{capital}, b) \rightarrow (b, \text{partOf}, a)$
$(c, \text{partOf}, d), (d, \text{partOf}, e) \rightarrow (c, \text{partOf}, e)$

OUTPUT: $\{(x \mapsto \text{Ireland}, y \mapsto \text{Europe}), (x \mapsto \text{Dublin}, y \mapsto \text{Ireland})$
$(x \mapsto \text{Dublin}, y \mapsto \text{Europe})\}$

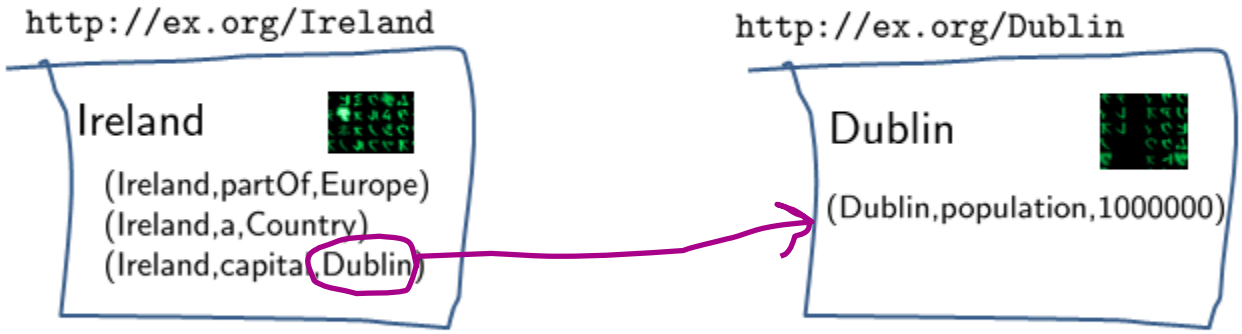
But what are we missing from here before we can have a "Web of Data"?

We need links!

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

INPUT: "(x,partOf,y)"

DATA:



RULES: $(a, \text{capital}, b) \rightarrow (b, \text{partOf}, a)$
$(c, \text{partOf}, d), (d, \text{partOf}, e) \rightarrow (c, \text{partOf}, e)$

OUTPUT: $\{(x \mapsto \text{Ireland}, y \mapsto \text{Europe}), (x \mapsto \text{Dublin}, y \mapsto \text{Ireland})$
$(x \mapsto \text{Dublin}, y \mapsto \text{Europe})\}$

… can't have a Web without links!

We need links!

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

INPUT: "(x,partOf,y)"

DATA:

http://ex.org/Ireland

Ireland 

(Ireland,partOf,Europe)
(Ireland,a,Country)
(Ireland,capital,Dublin)

http://ex.org/Dublin

Dublin 

(Dublin,population,1000000)

RULES: $(a, \text{capital}, b) \rightarrow (b, \text{partOf}, a)$
$(c, \text{partOf}, d), (d, \text{partOf}, e) \rightarrow (c, \text{partOf}, e)$

OUTPUT: $\{(x \mapsto \text{Ireland}, y \mapsto \text{Europe}), (x \mapsto \text{Dublin}, y \mapsto \text{Ireland})$
$(x \mapsto \text{Dublin}, y \mapsto \text{Europe})\}$

So how can we include links in this picture?

RDF filled with IRIs!

```
@base <http://ex1.org/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix ex1: <http://ex1.org/#> .
<#Jen> a <http://ex1.org/#Person> , ex1:Female ;
   rdfs:label "Jen"@en ; <#allergy> <#Citrus> ;
   ex1:location [ ex1:lat 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 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 <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 <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> .
```

```
ame "Álvarez"@es;
hday "02-24";
er "female";
page <http://anakena.dcc.uchile.cl/~calvarez/cv.html>;
d_near <http://sws.geonames.org/3871336/>;
erest <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> .
```

HTTP GET

```
@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#> .

<#me> 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
;
 foaf:nick "Ritto" ;
 foaf:birthday "25/08" .
```

```
ame "Álvarez"@es;
hday "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>;
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> .
```

HTTP GET

```
@prefix dbo:    <http://dbpedia.org/ontology/> .
@prefix ns1:    <http://dbpedia.org/resource/Tales_of_Symphonia:> .
ns1:_Dawn_of_the_New_World      dbo:series      <http://dbpedia.org/resource/Tales_(series)> .
@prefix dbr:    <http://dbpedia.org/resource/> .
dbr:Yoshito     dbo:wikiPageDisambiguates       <http://dbpedia.org/resource/Tales_(series)> .
@prefix foaf:   <http://xmlns.com/foaf/0.1/> .
<http://en.wikipedia.org/wiki/Tales_(series)>   foaf:primaryTopic
<http://dbpedia.org/resource/Tales_(series)> .
@prefix ns4:    <http://dbpedia.org/resource/Tales_of_the_Heroes:> .
ns4:_Twin_Brave dbo:series      <http://dbpedia.org/resource/Tales_(series)> .
@prefix ns5:    <http://dbpedia.org/resource/Keroro_RPG:> .
ns5:_Kishi_to_Musha_to_Densetsu_no_Kaizoku      dbo:series
<http://dbpedia.org/resource/Tales_(series)> .
@prefix dbp:    <http://dbpedia.org/property/> .
dbr:Bandai_Namco_Games  dbp:products    <http://dbpedia.org/resource/Tales_(series)> ;
        dbo:product     <http://dbpedia.org/resource/Tales_(series)> .
dbr:Tales_of_Symphonia_Chronicles       dbo:series      <http://dbpedia.org/resource/Tales_(series)> .
dbr:Tales_of_Hearts     dbo:series      <http://dbpedia.org/resource/Tales_(series)> .
dbr:Tales_of_Graces     dbo:series      <http://dbpedia.org/resource/Tales_(series)> .
dbr:Tales_of_Vesperia   dbo:series      <http://dbpedia.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 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 <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 <http://sws.geonames.org/3871336/>;
    foaf:interest <http://dbpedia.org/resource/Tales_(series)>,
p://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>,
p://anakena.dcc.uchile.cl/~jogarrid/web/foaf.ttl#yo>,
p://anakena.dcc.uchile.cl/~ekauffma/foaf.ttl#eli> .
```

HTTP GET

```
@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 …

Semantic Web, early days (pre-2006)

- Lots of dumps of RDF

Index of /pub/databases/uniprot/current_release/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

Semantic Web, early days (pre-2006)

- 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

- Classes by Name
- Properties by Name

Statistics

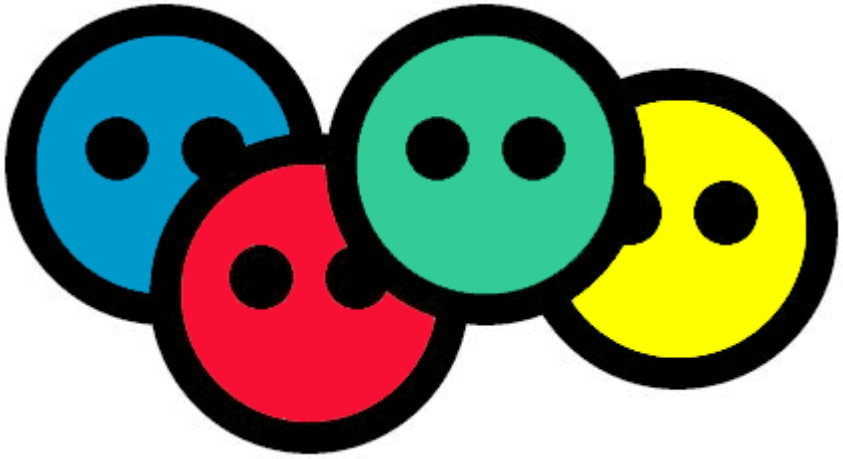
- 282 ontologies
- Number of Classes, Properties, and Instances Defined
- DAML Property (Feature) Use by Ontology

Semantic Web, early days (pre-2006)

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

Semantic Web, early days (pre-2006)

- 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

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

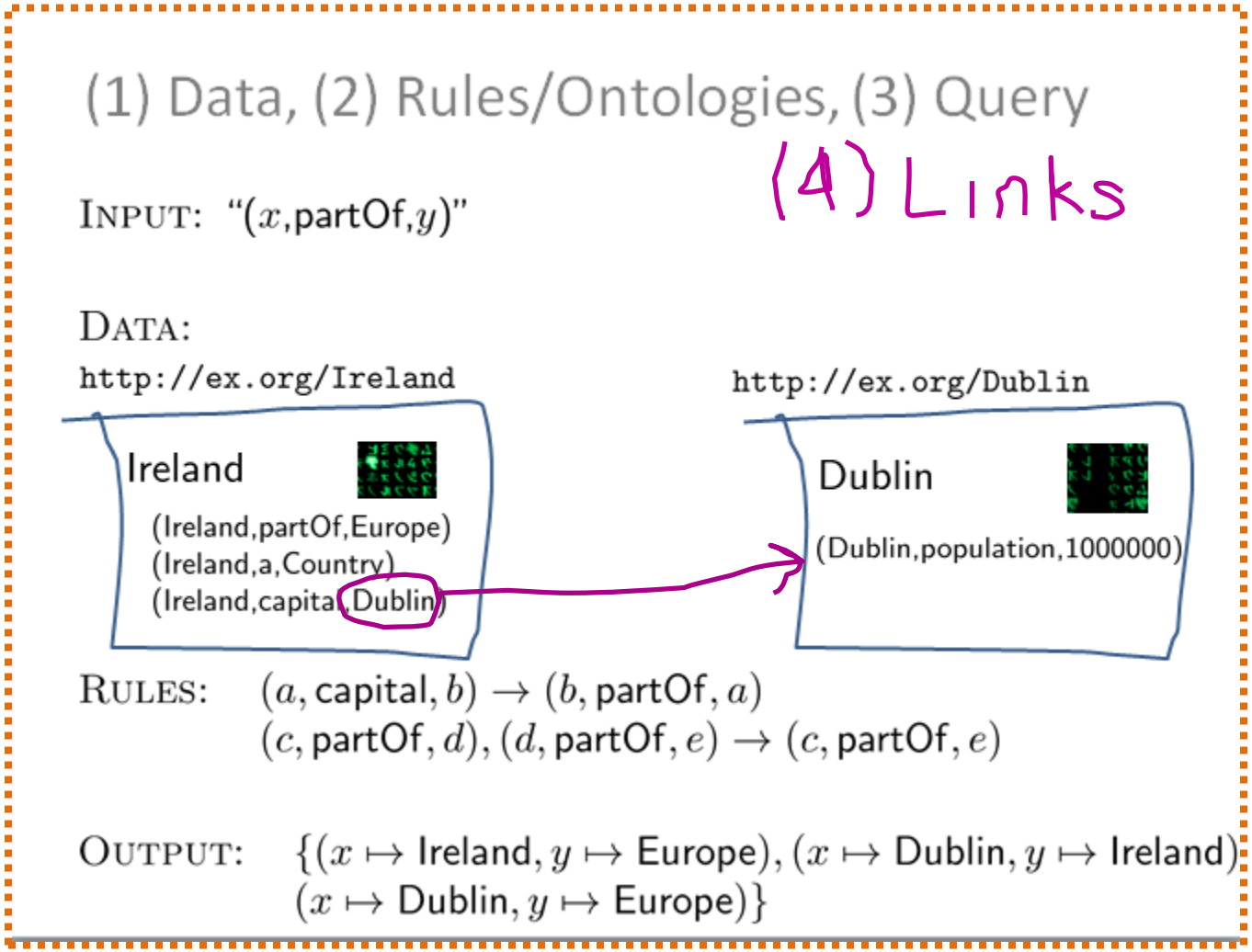
This is Linked Data …

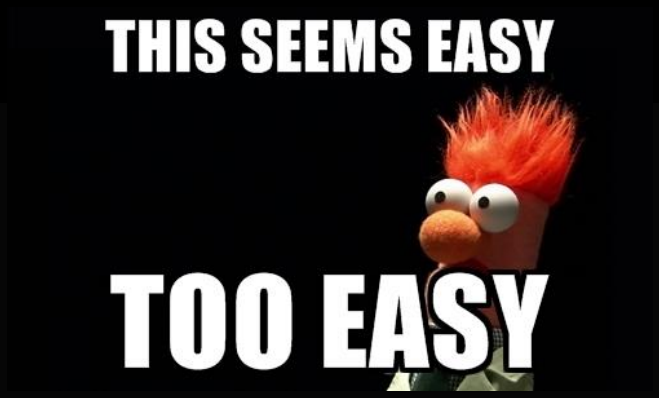
```
@base <http://anakena.dcc.uchile.cl/~calvarez/foaf.ttl> .

@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 <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> .
```

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`

Problems?

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

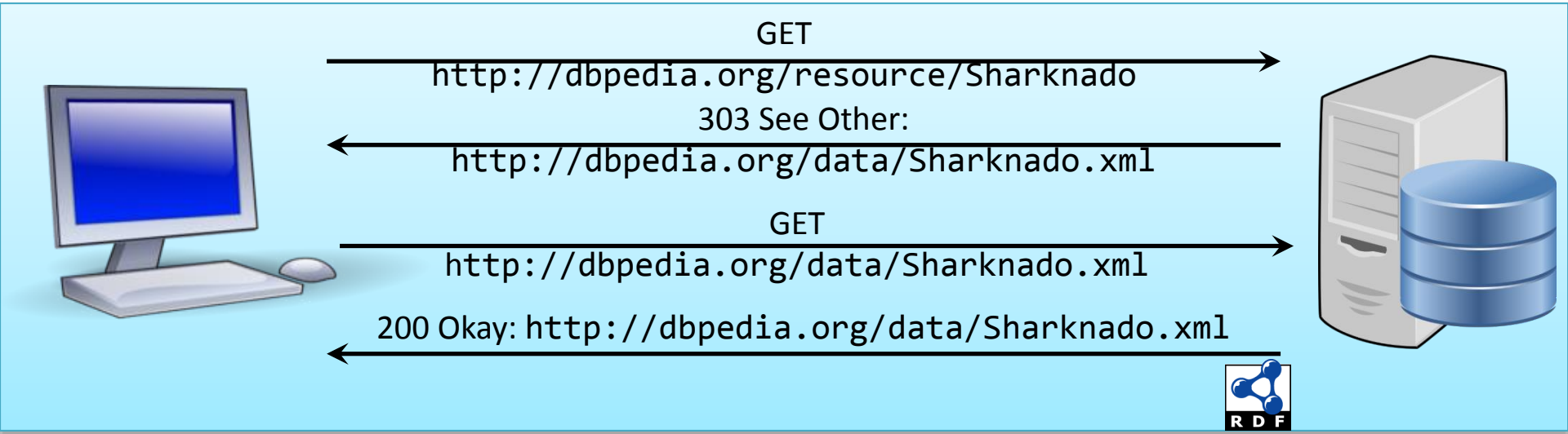
Solutions?

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

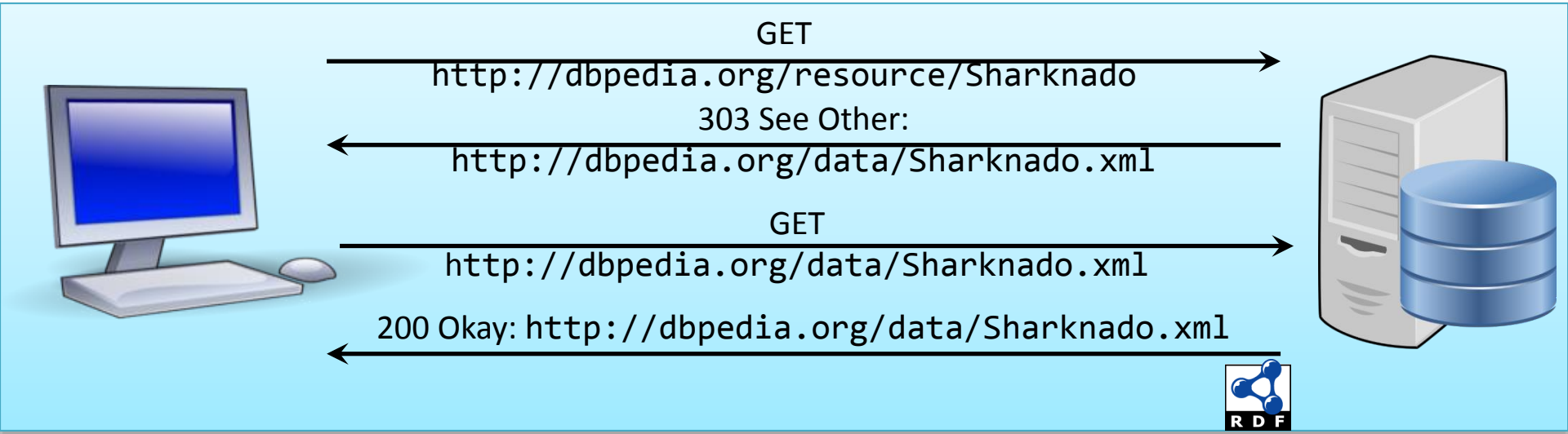
2. Slash (with `303 redirect`)
   - `http://dbpedia.org/data/Sharknado.xml`
     - Identifies RDF document about Sharknado
   - `http://dbpedia.org/resource/Sharknado`
     - Identifies the movie Sharknado
     - Look it up, it redirects with `303` to RDF doc. above



Hash vs. Slash

Which is better, hash or slash?

Well, hash has half the number of requests!



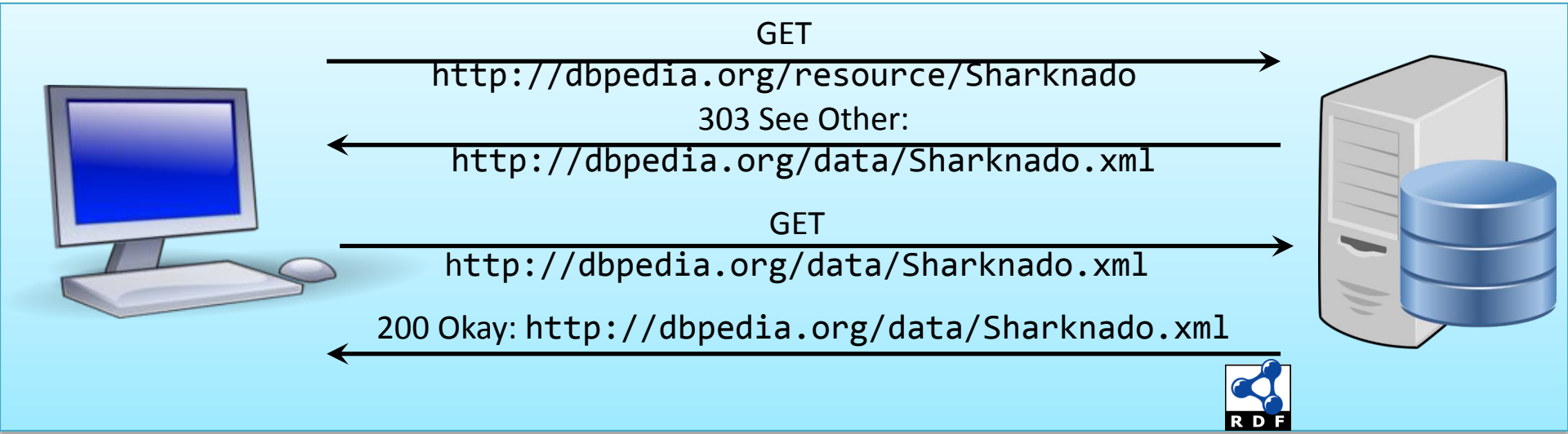
Why does slash exist then?

- Hash often used to describe multiple resources in one document

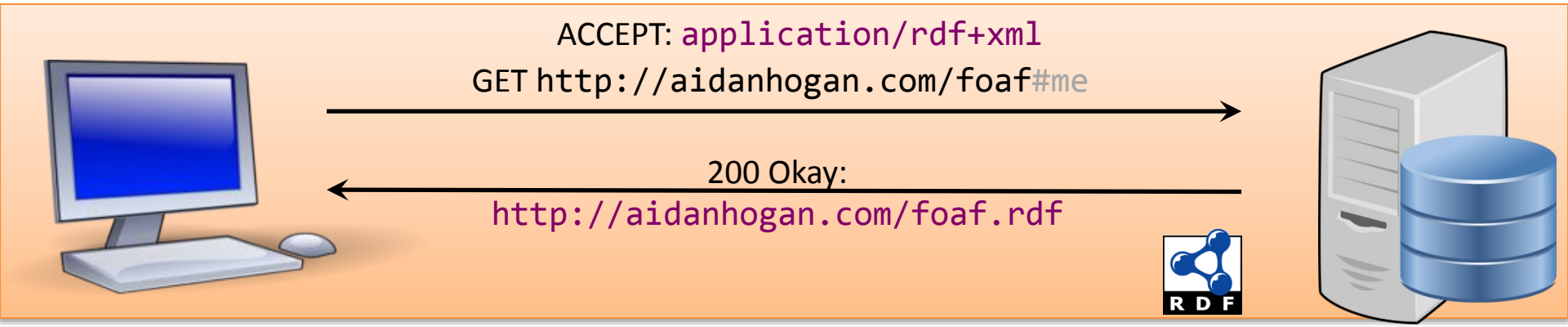
- Slash often used when there are too many resources to fit in one document
  - But nothing would stop, e.g.:
    - `http://dbpedia.org/data/Sharknado.xml#this`

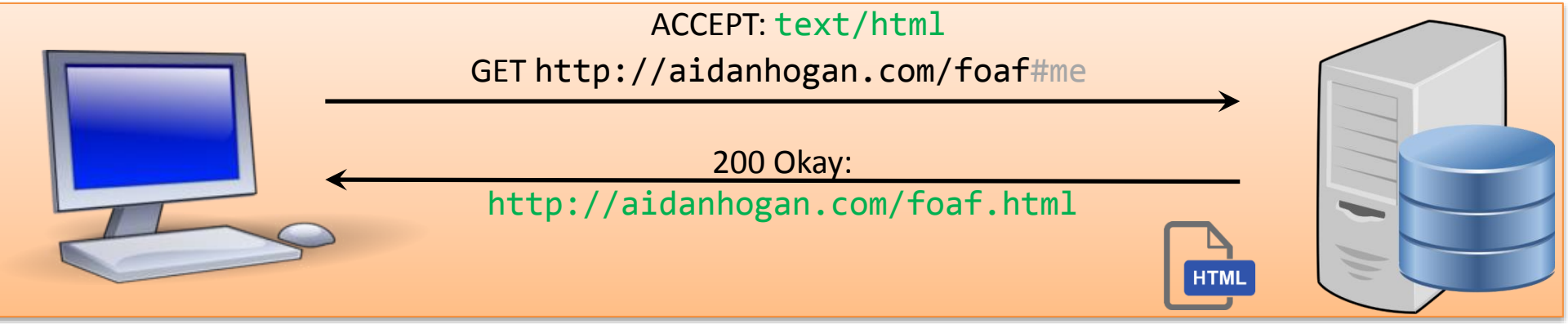
So … erm … why does slash exist then?

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



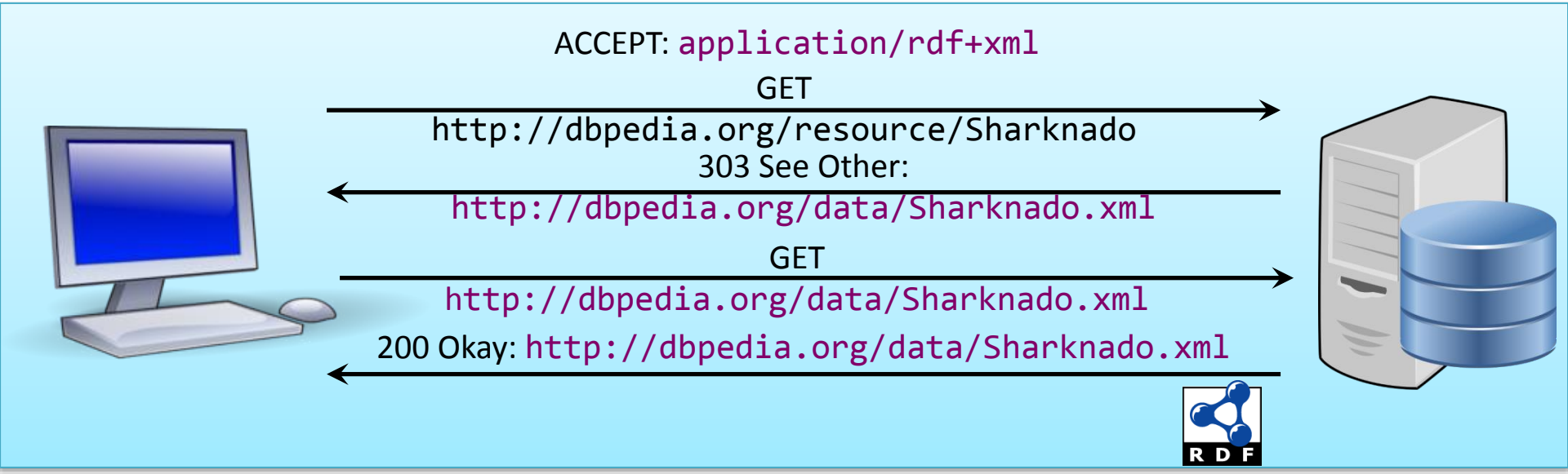
Content negotiation with hash

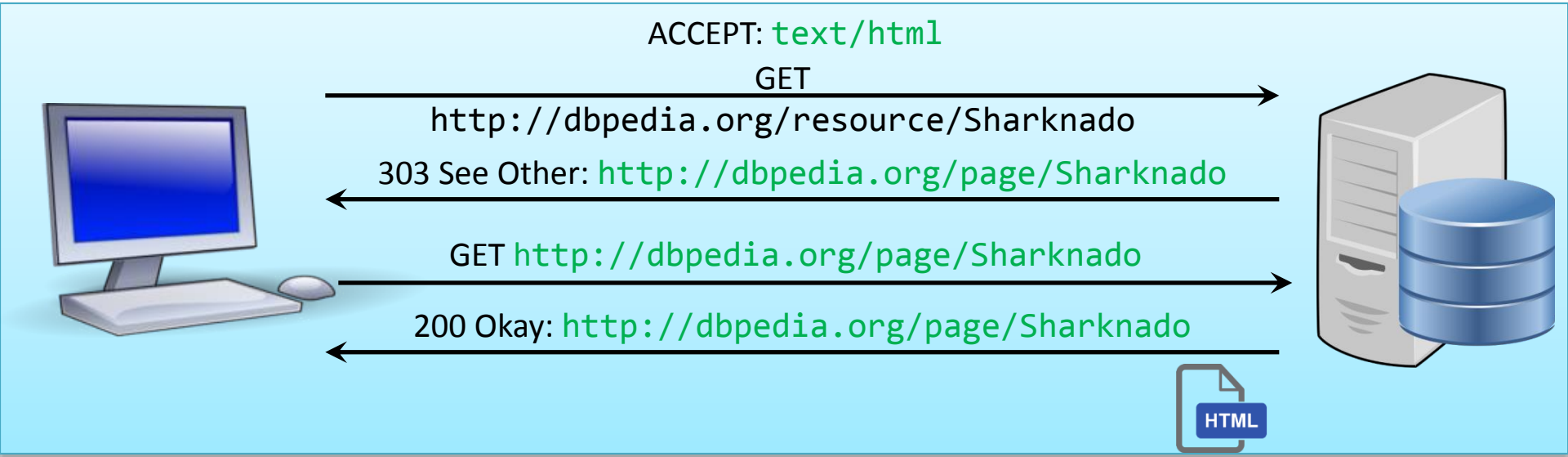




Can also choose from different RDF formats; e.g., Turtle, RDFa, etc.
(if supported by the server that is!)
Can also specify preferences using q values …

Content negotiation with slash





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

We've got all these people who want to publish Open Data but how should they publish it on the Web?

We've got this new way of publishing Linked Data on the Web but no data to publish …

Linked Open Data

Are you thinking what I'm thinking?

The 5 ★'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

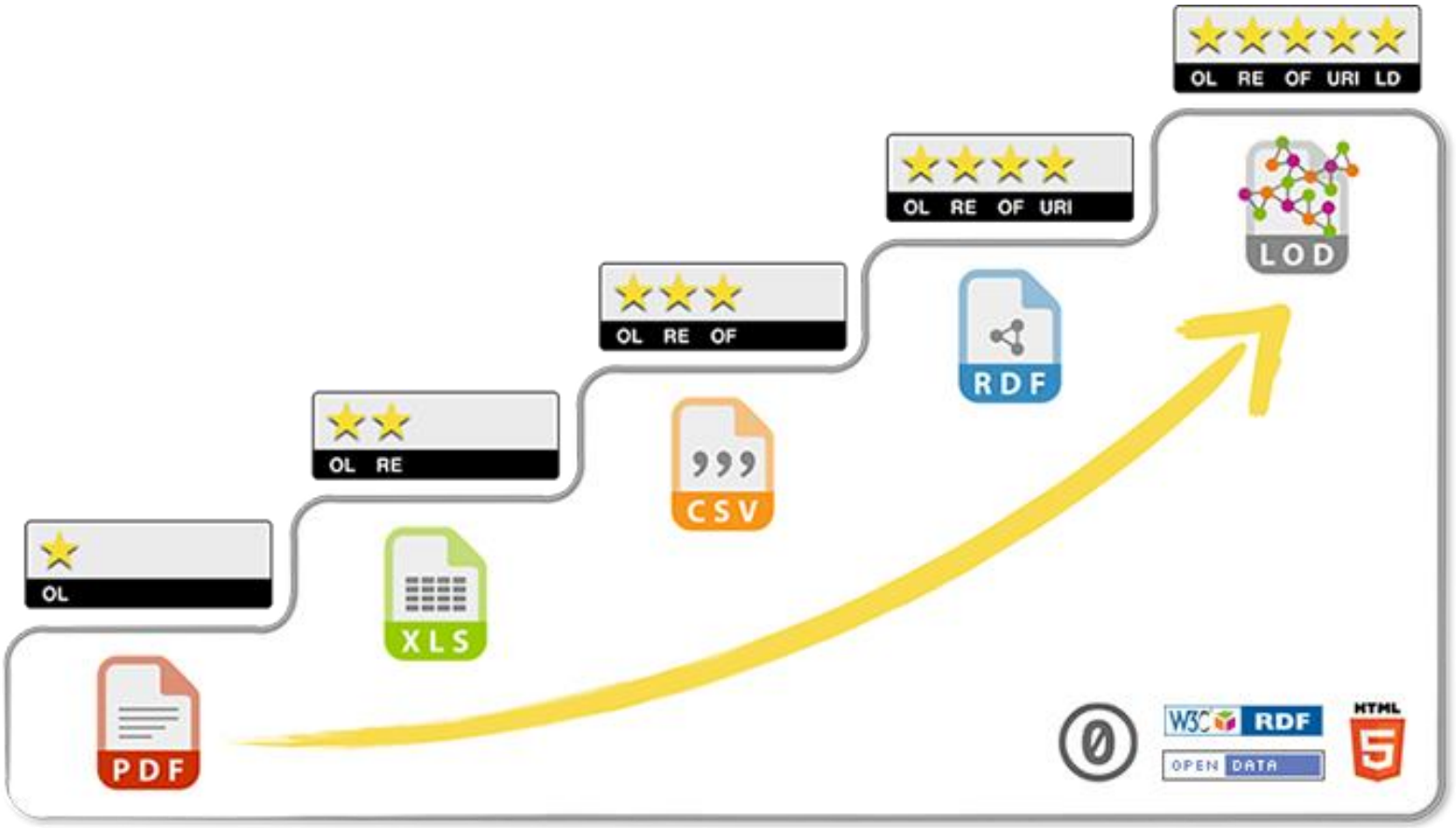
★★★★ Use URIs to name your stuff (hint: RDF)

- use unambiguous identifiers that can be linked/looked up

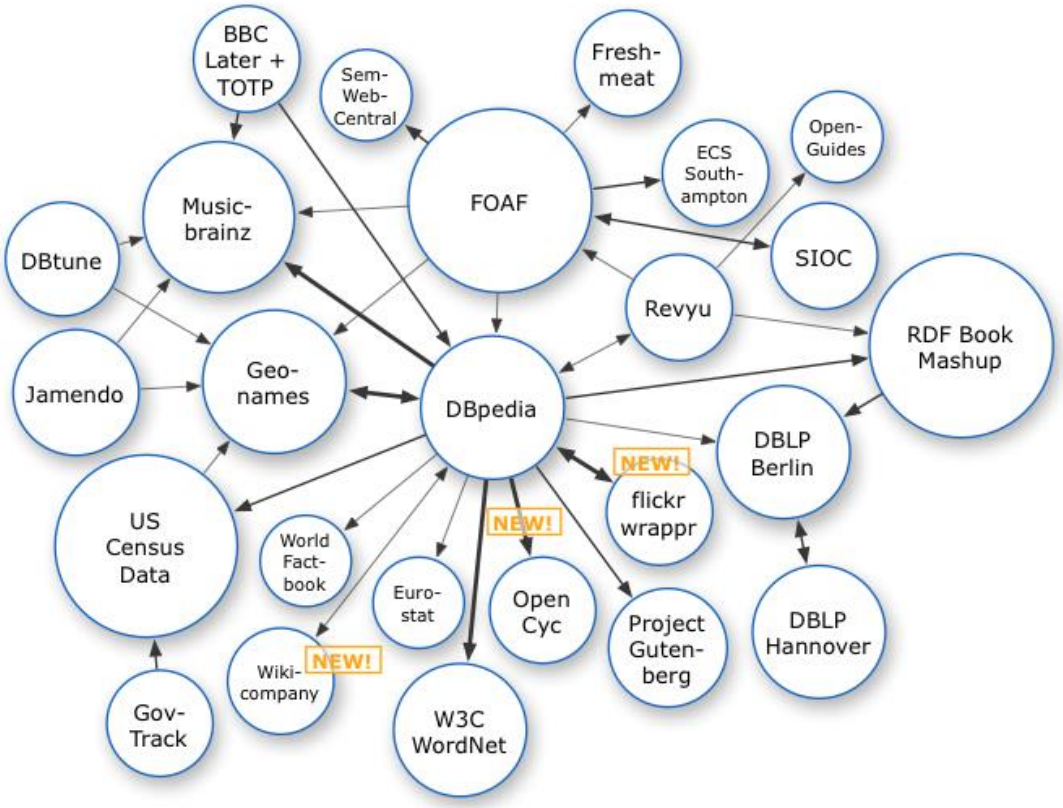
★★★★★ Provide links to other content

- so consumers can follow links to find out more

Each star improves interoperability of data

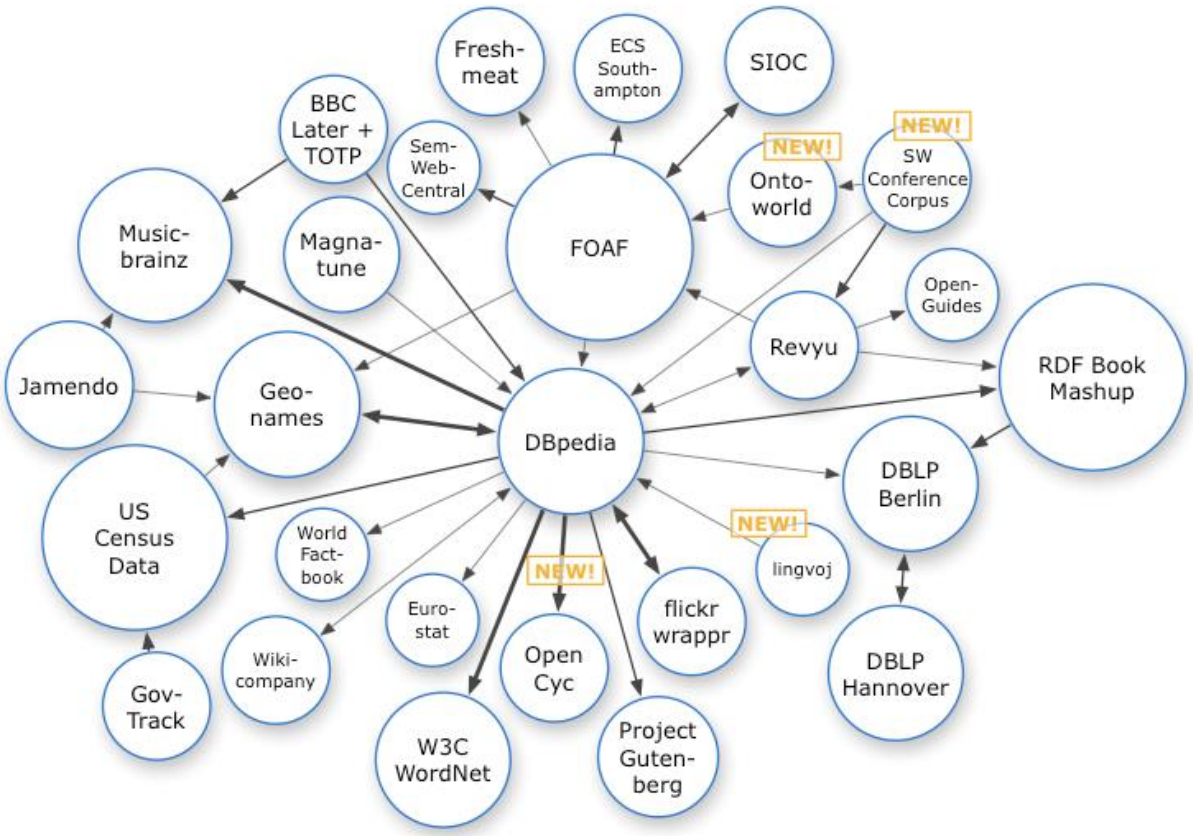


Growth of the Linked Data Cloud



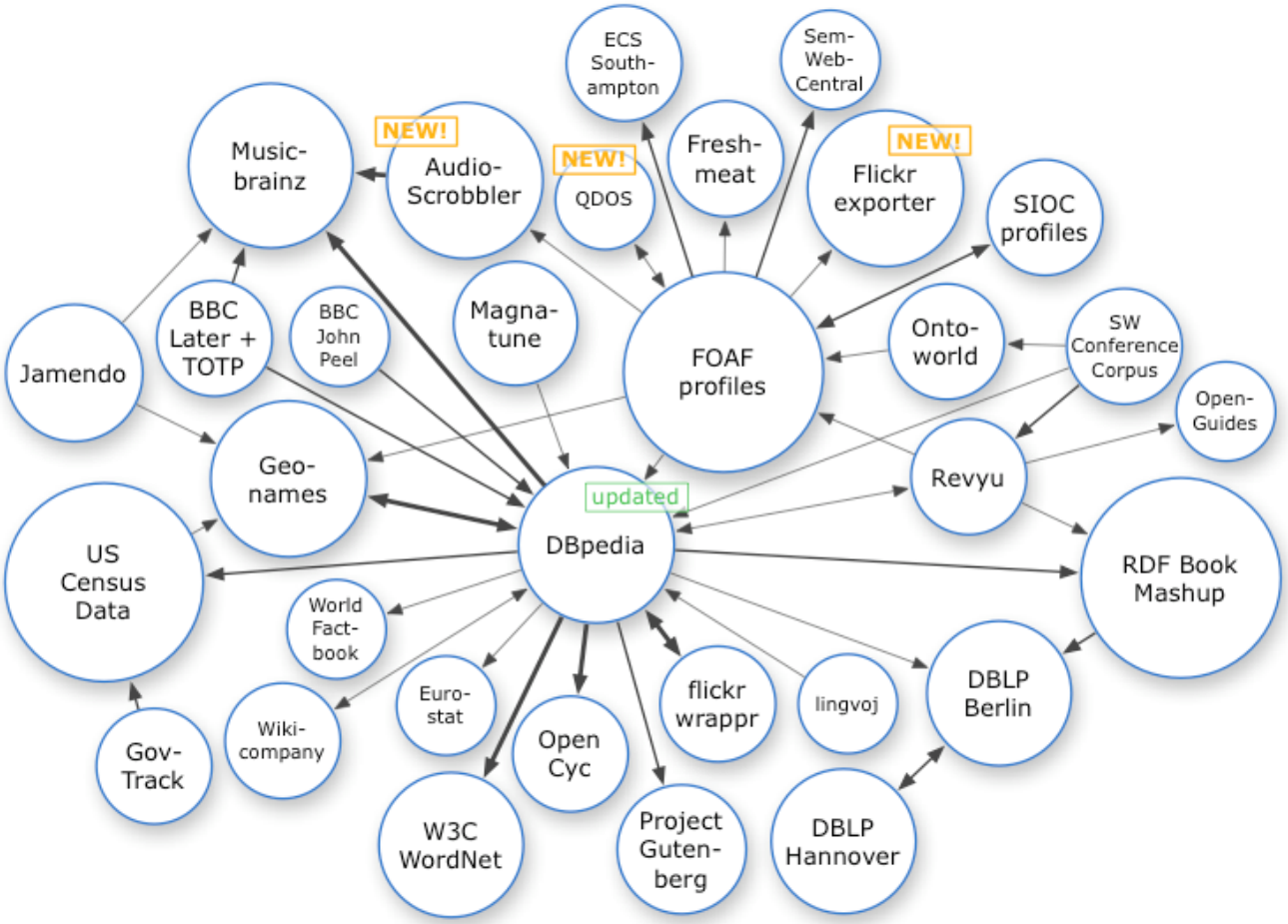
Oct. 2007

Growth of the Linked Data Cloud



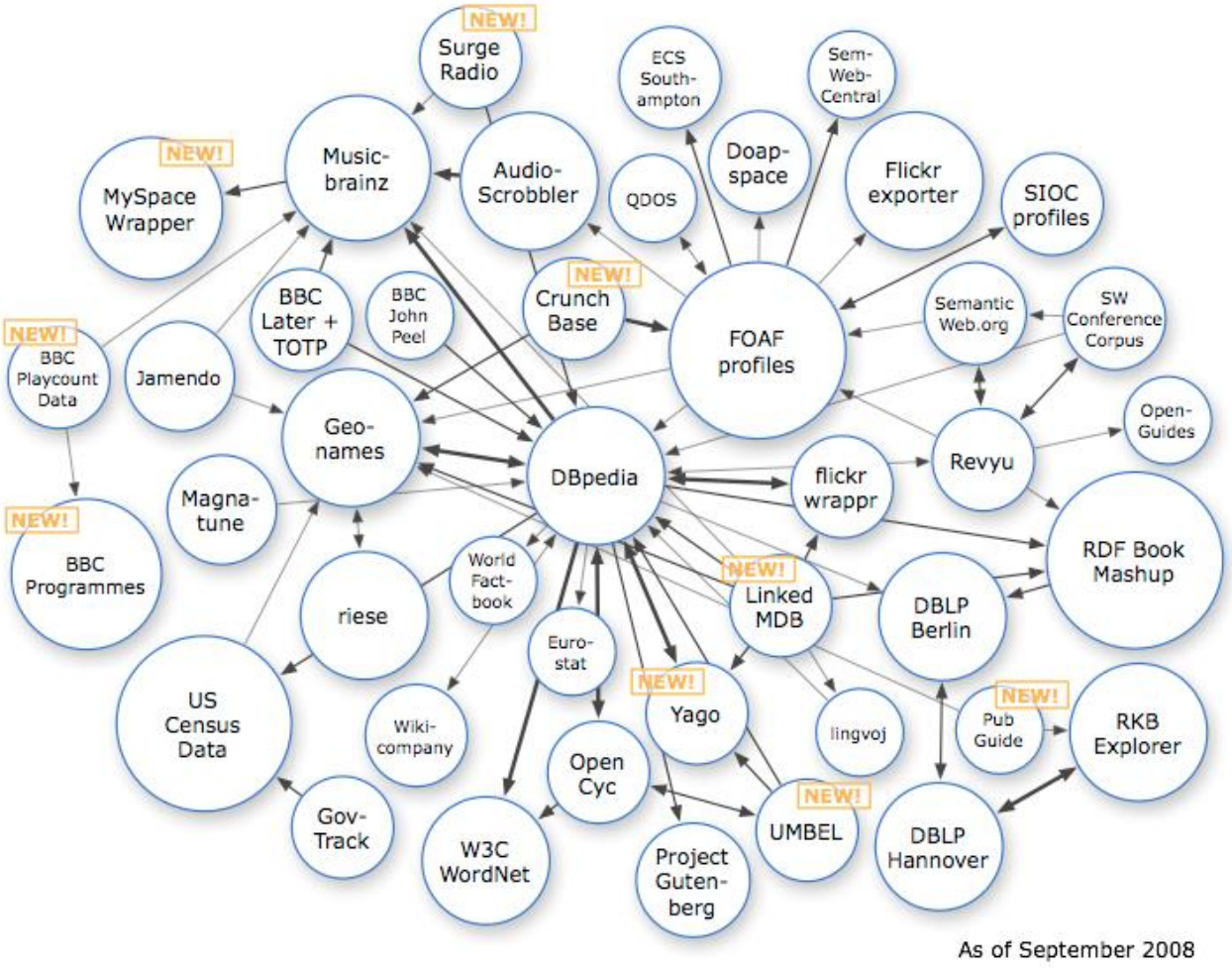
Oct. 2007
Nov. 2007

Growth of the Linked Data Cloud



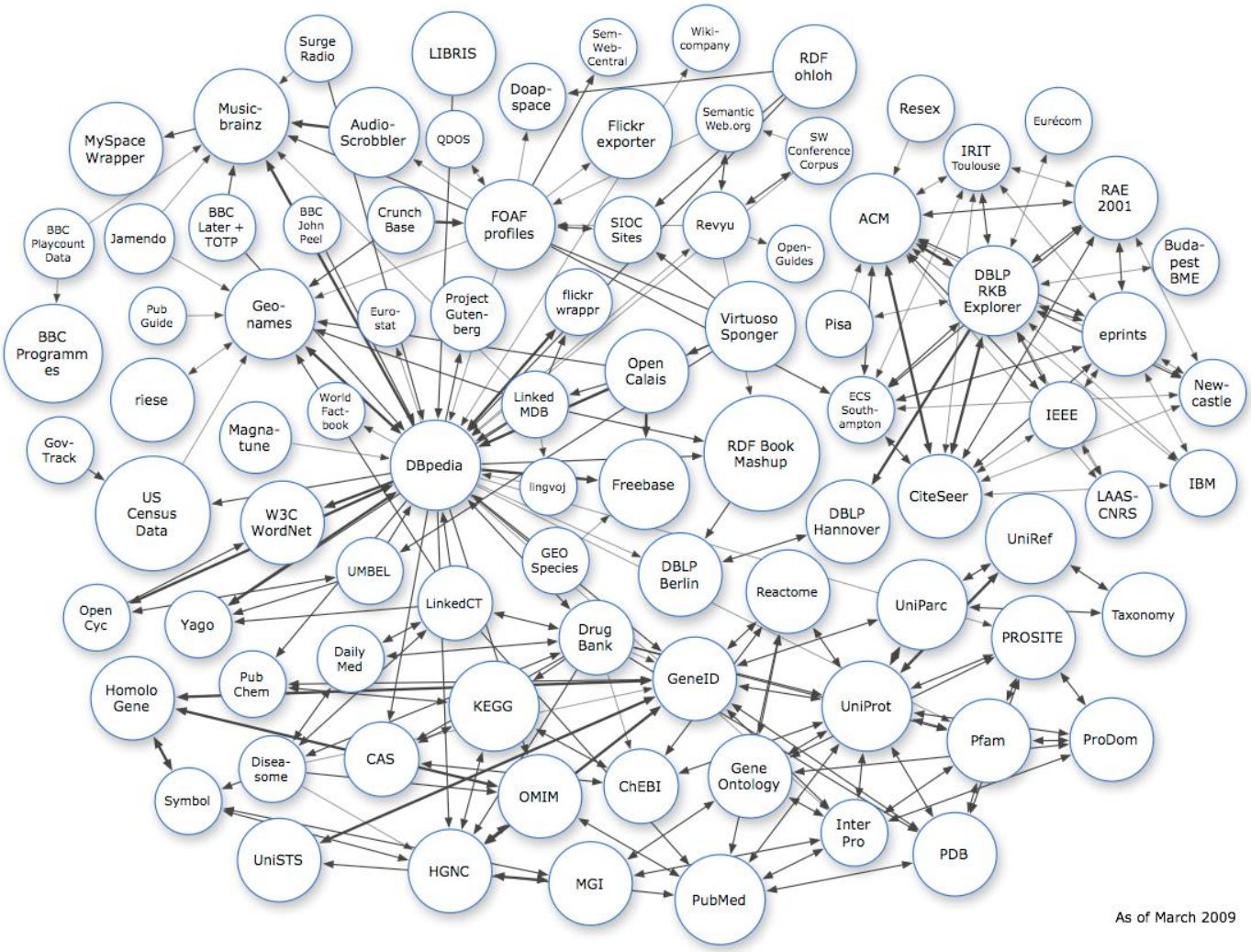
Oct. 2007
Nov. 2007
Feb. 2008

Growth of the Linked Data Cloud



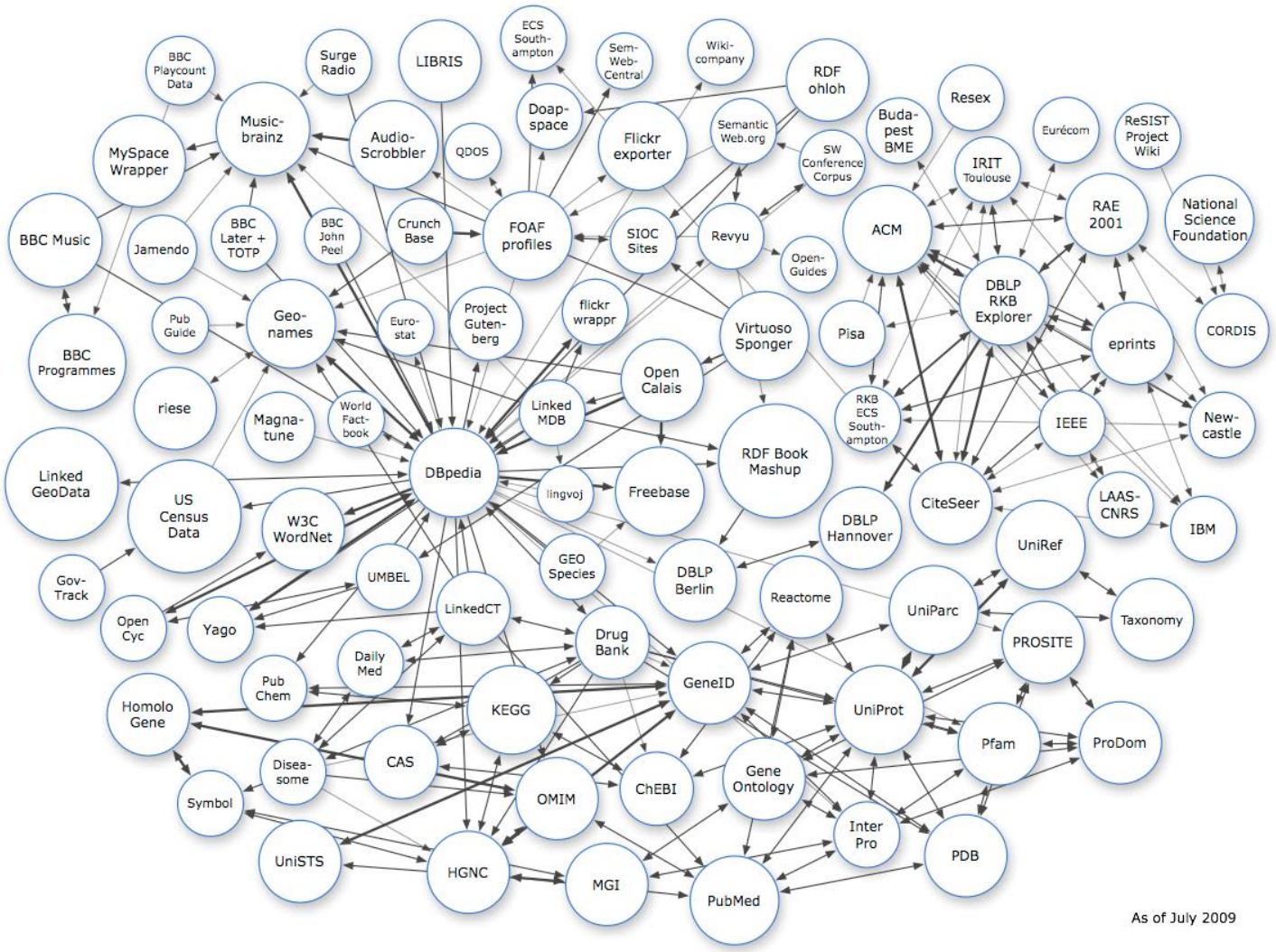
Oct. 2007
Nov. 2007
Feb. 2008
Sep. 2008

Growth of the Linked Data Cloud



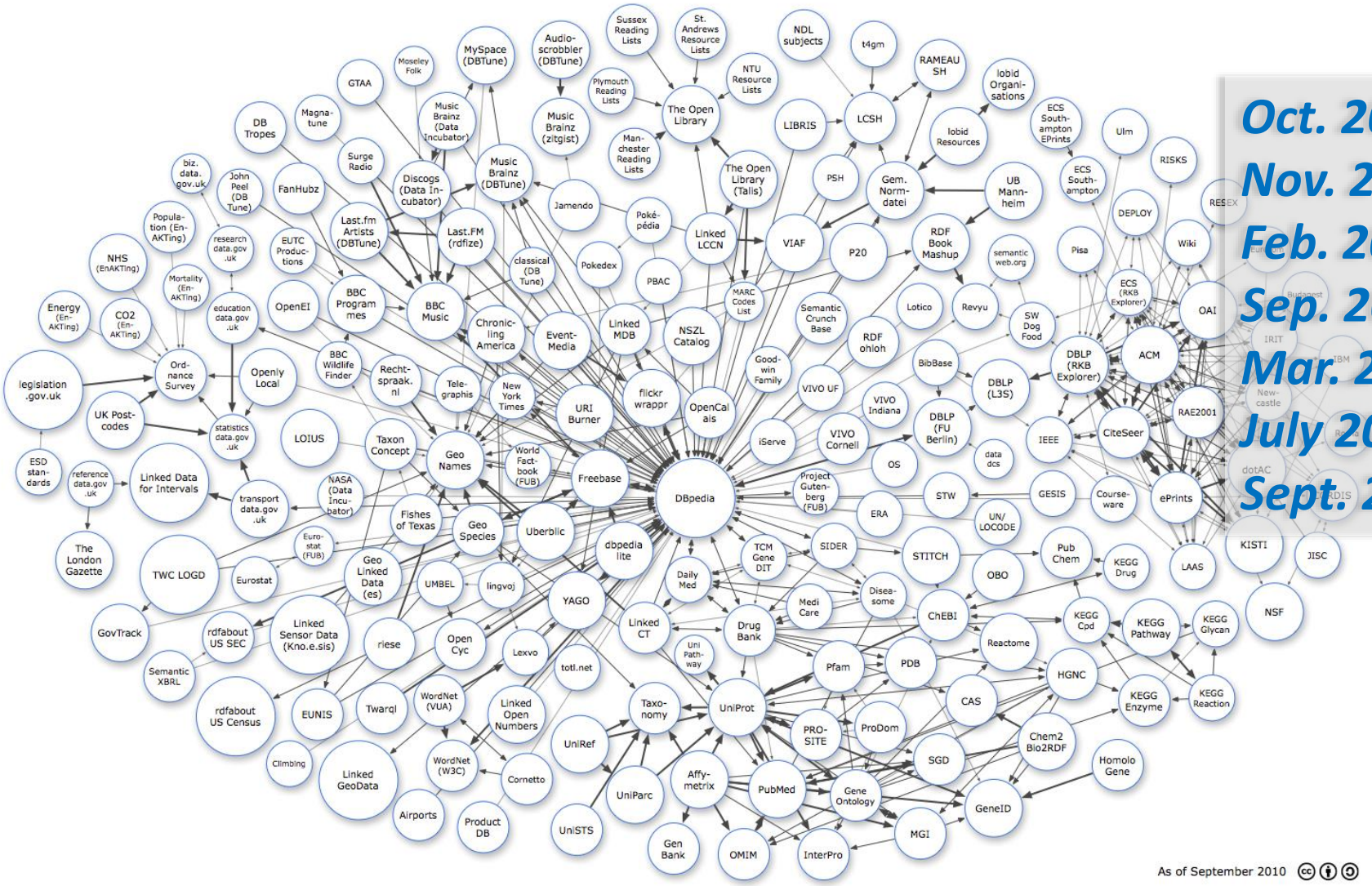
Oct. 2007
Nov. 2007
Feb. 2008
Sep. 2008
Mar. 2009

Growth of the Linked Data Cloud

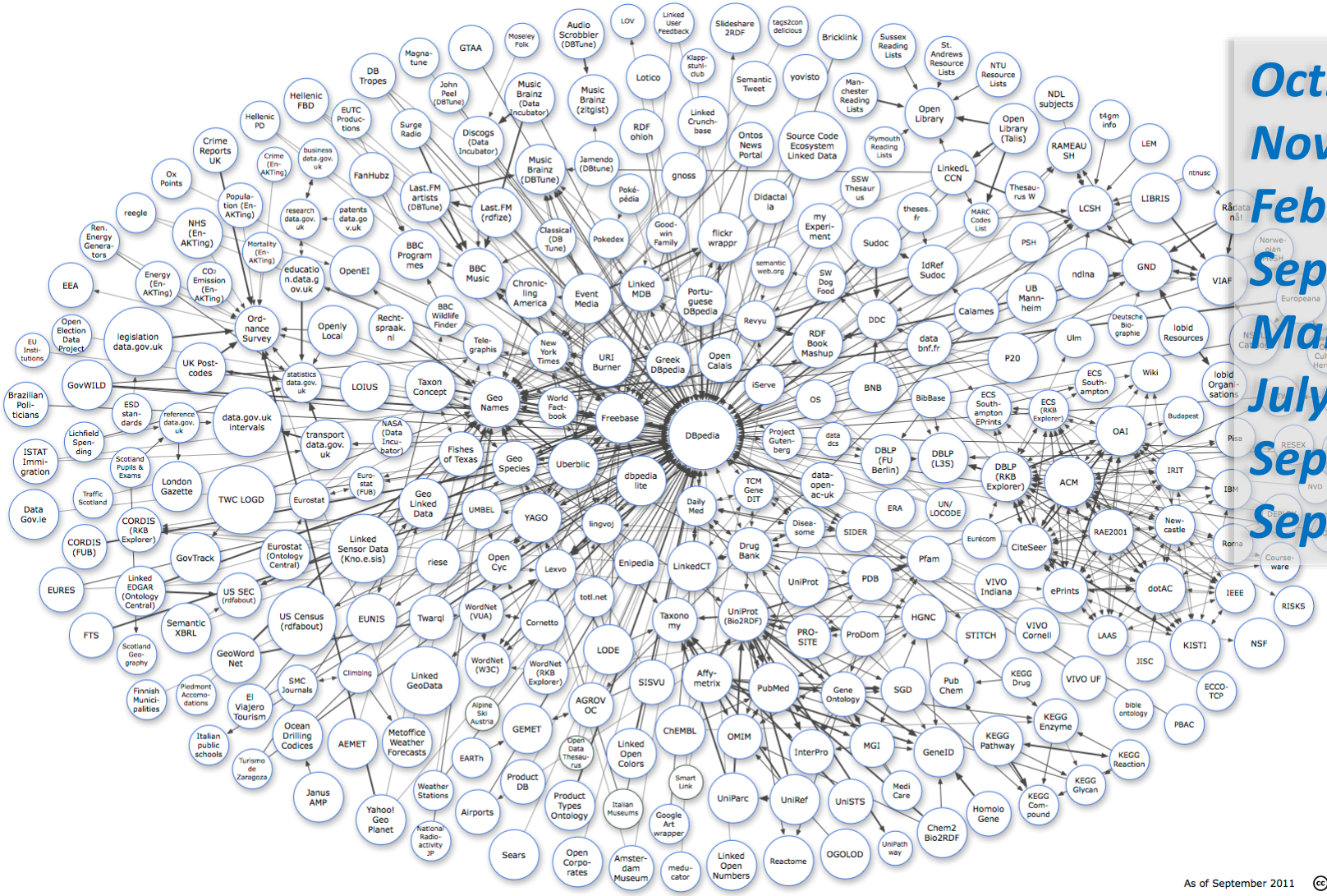


Oct. 2007
Nov. 2007
Feb. 2008
Sep. 2008
Mar. 2009
July 2009

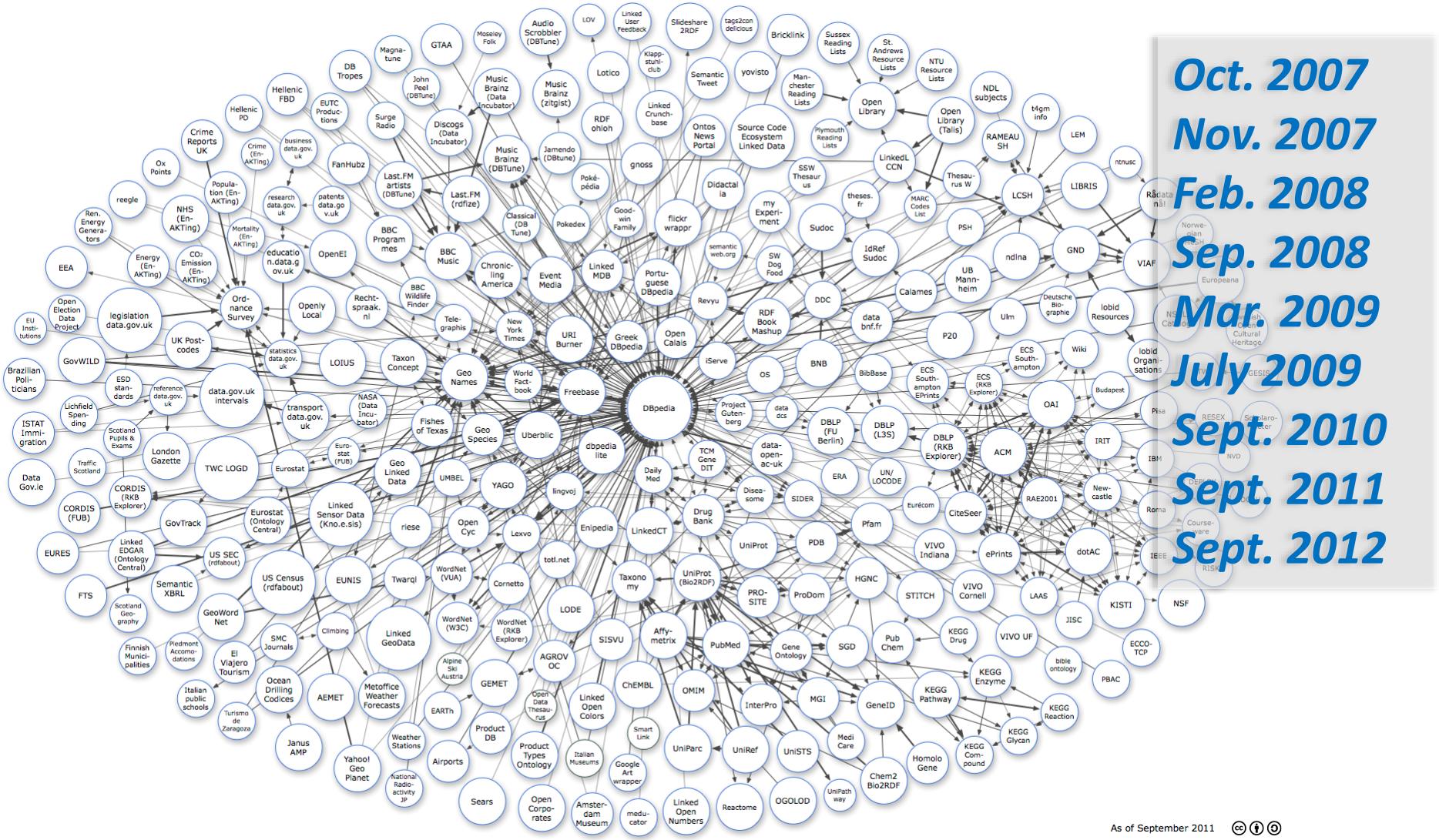
Growth of the Linked Data Cloud



Growth of the Linked Data Cloud

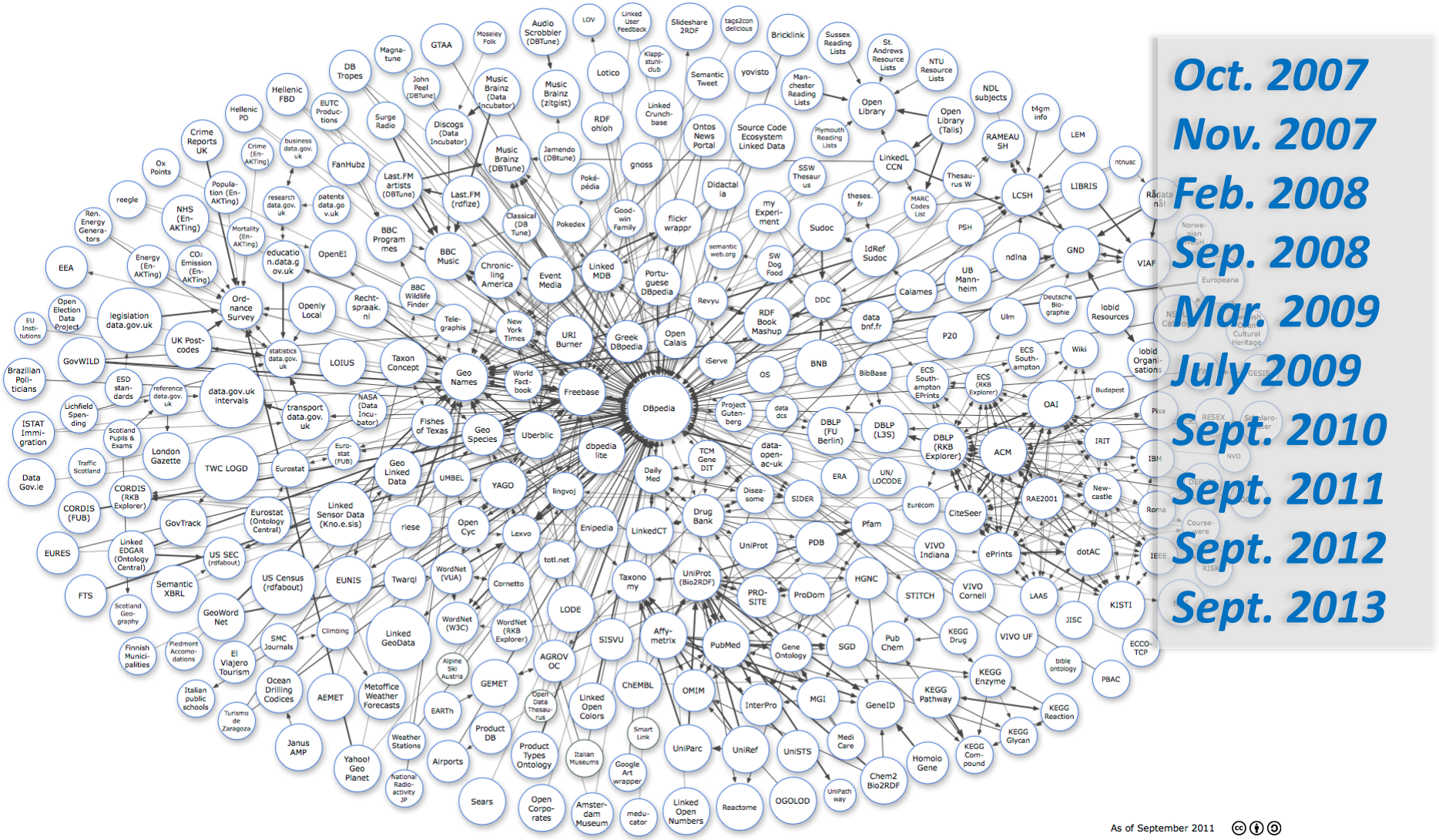


Growth of the Linked Data Cloud



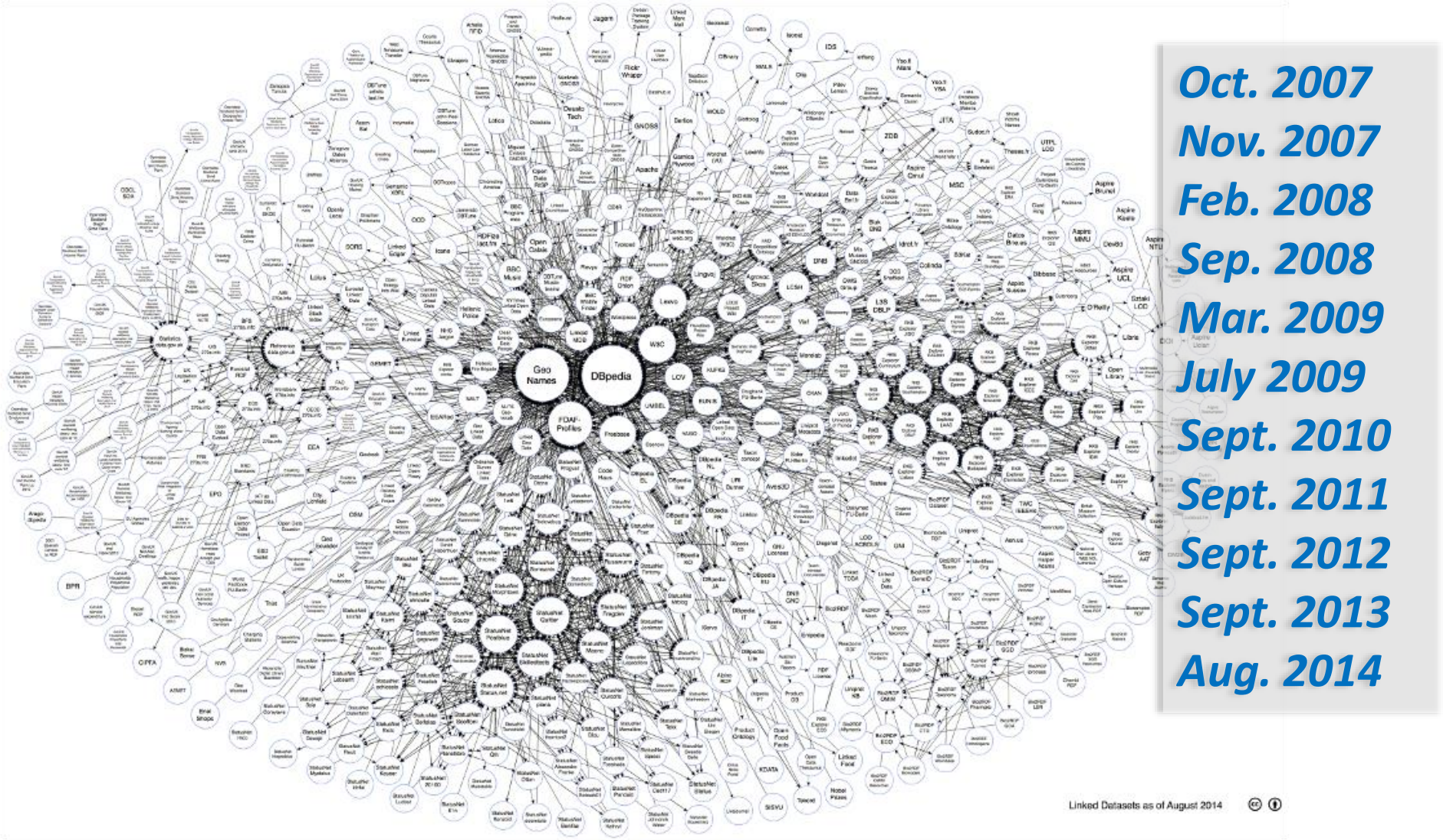
As of September 2011

Growth of the Linked Data Cloud



As of September 2011

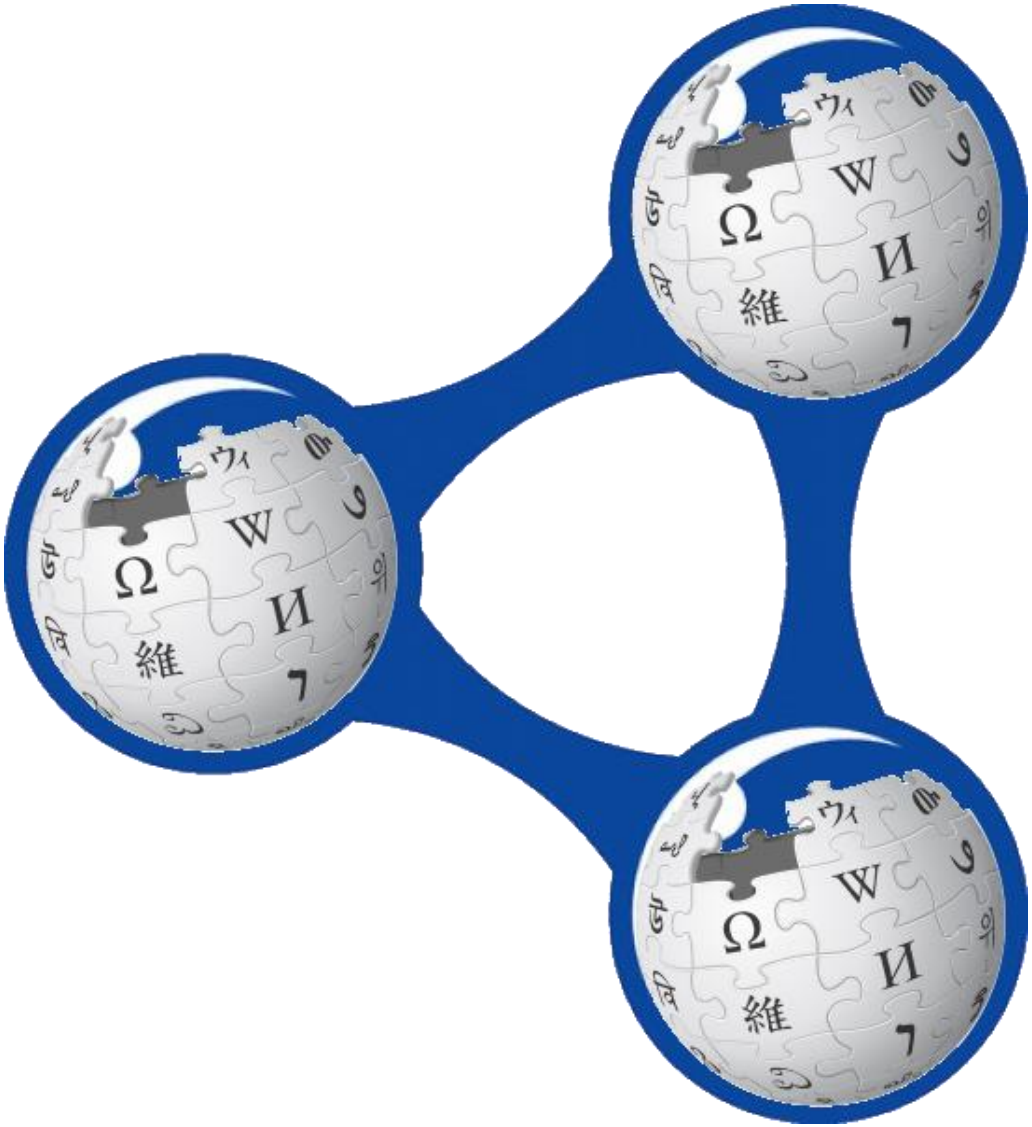
Growth of the Linked Data Cloud (sort of)



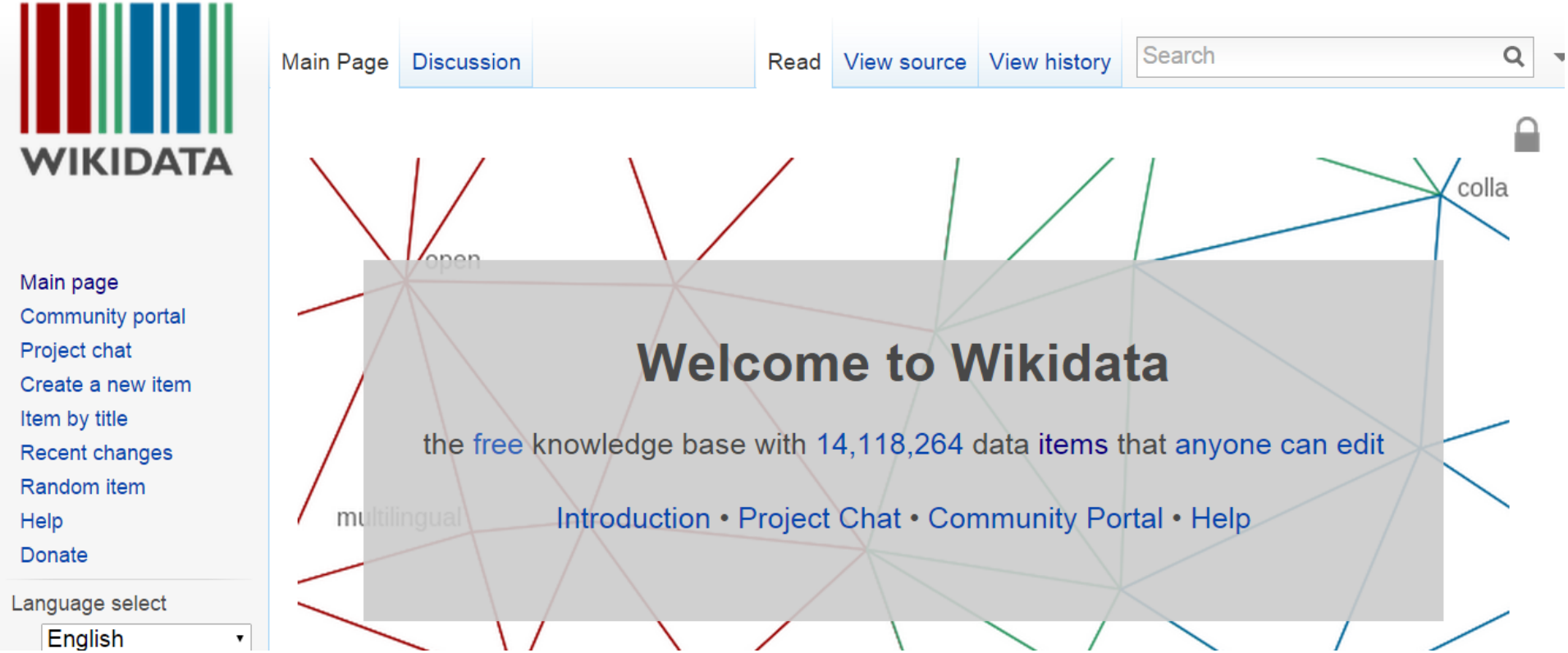
Oct. 2007
Nov. 2007
Feb. 2008
Sep. 2008
Mar. 2009
July 2009
Sept. 2010
Sept. 2011
Sept. 2012
Sept. 2013
Aug. 2014

THERE IS A LOT OF LINKED DATA OUT THERE NOW …

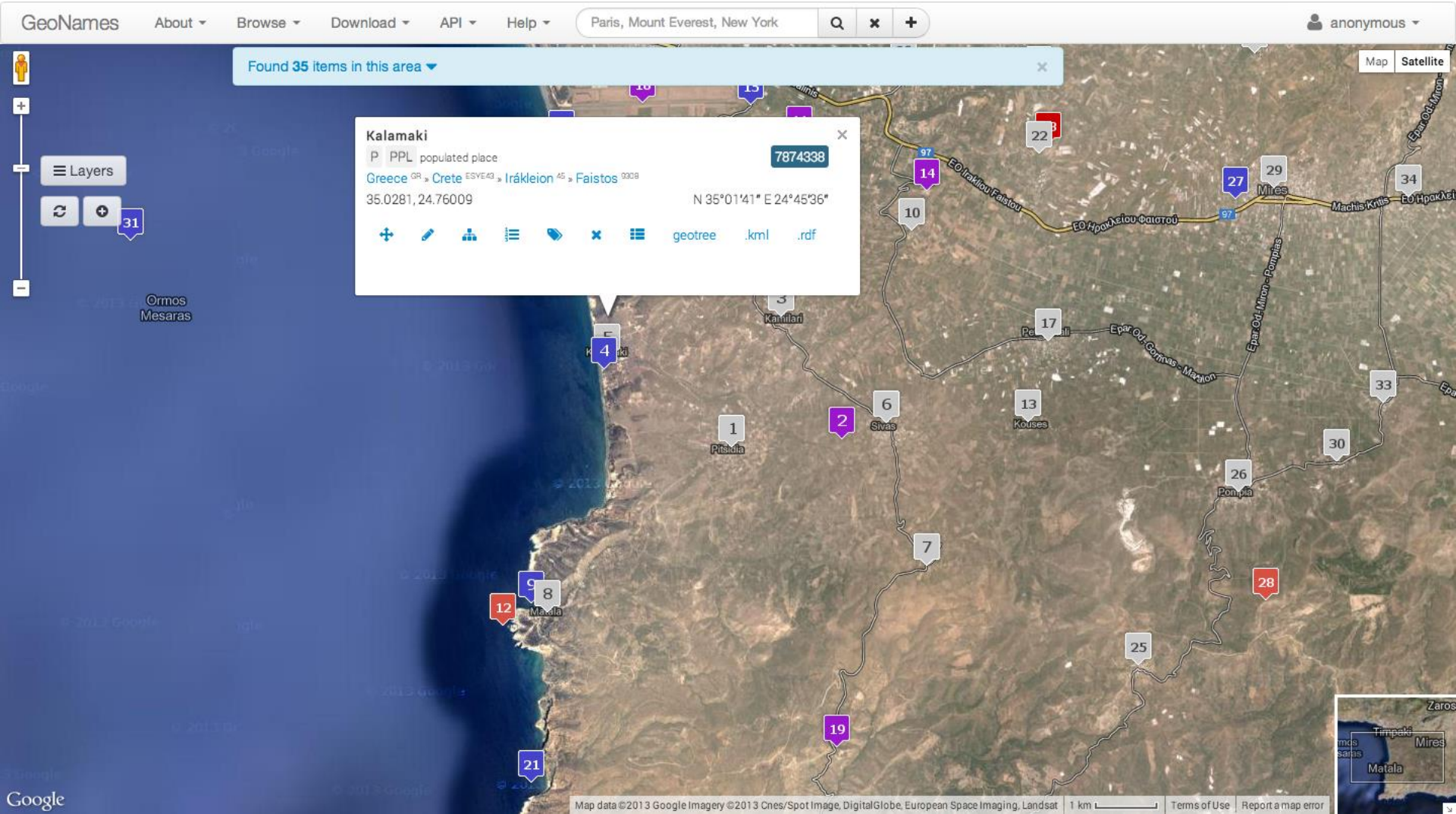
DBpedia:
An RDF Encyclopaedia



Wikidata:
Another RDF Encyclopaedia

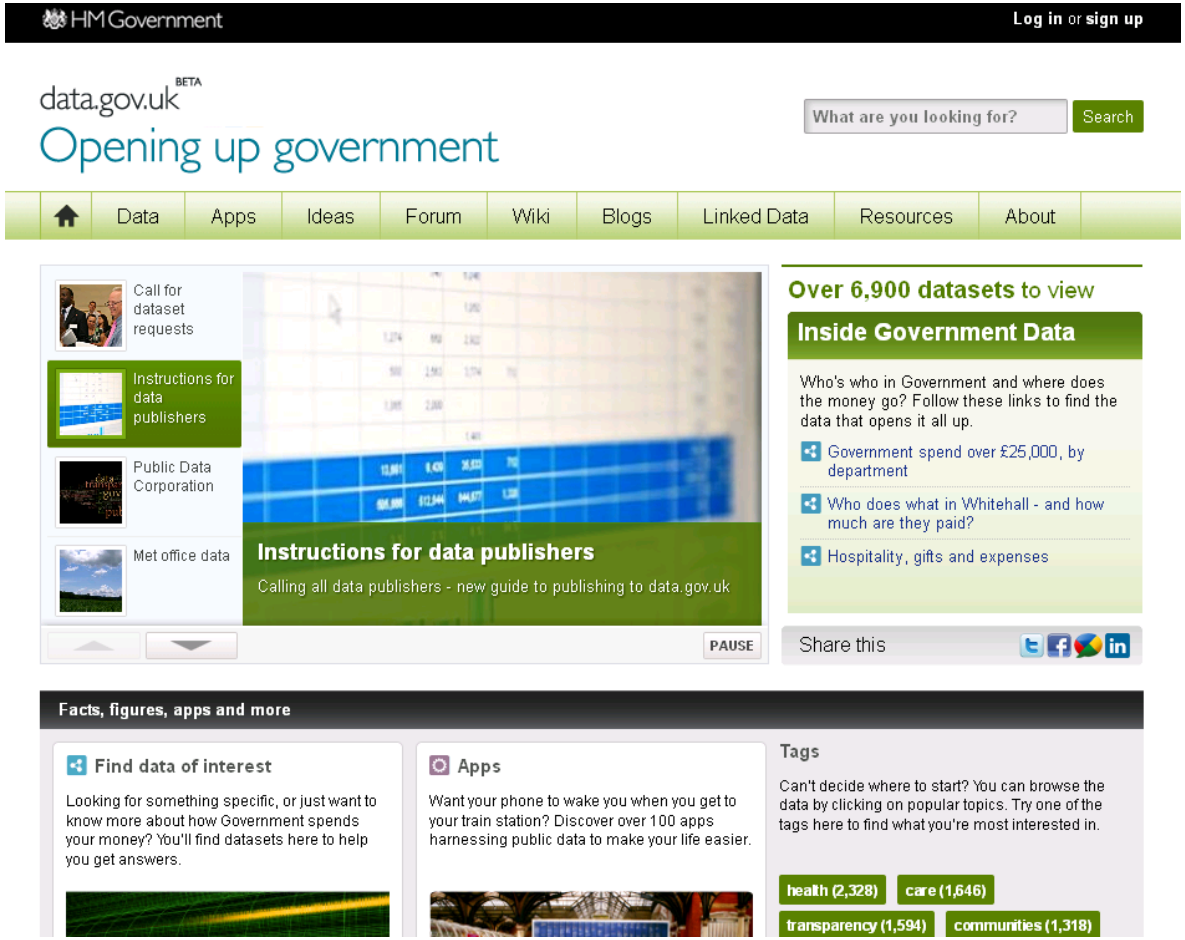


Geographical Data: Geonames

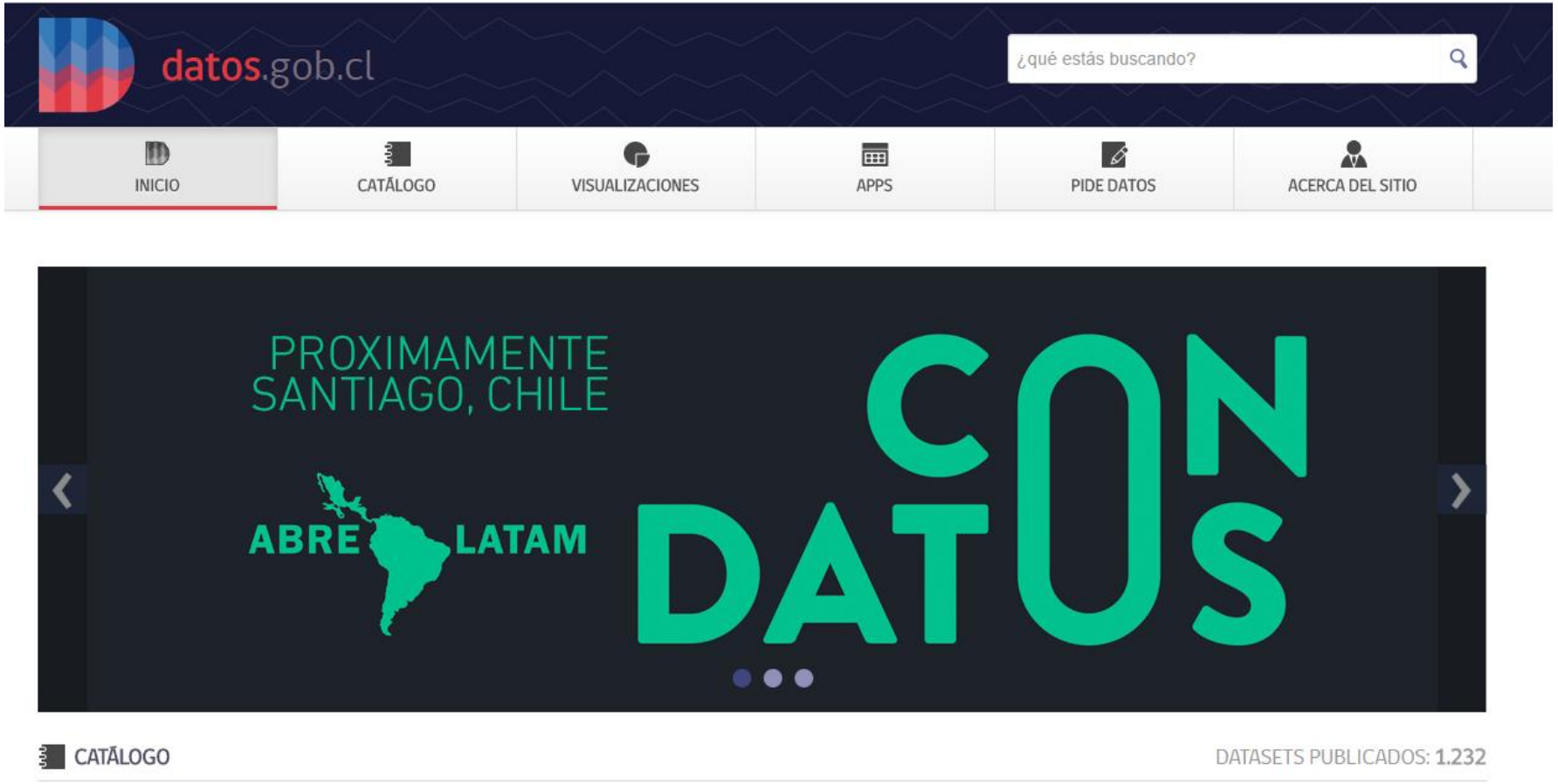


Linked Government Data: data.gov

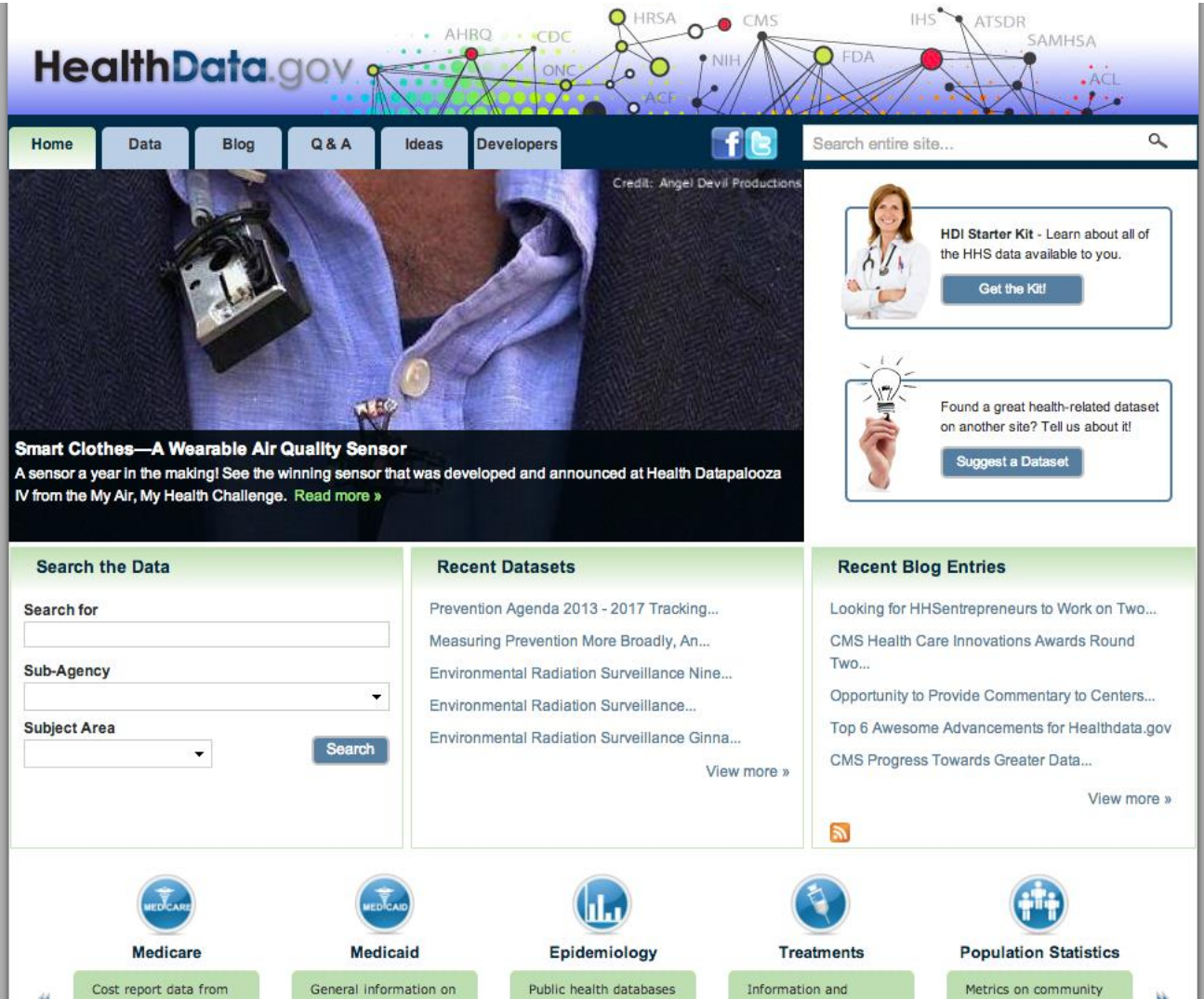
Linked Government Data: data.gov.uk



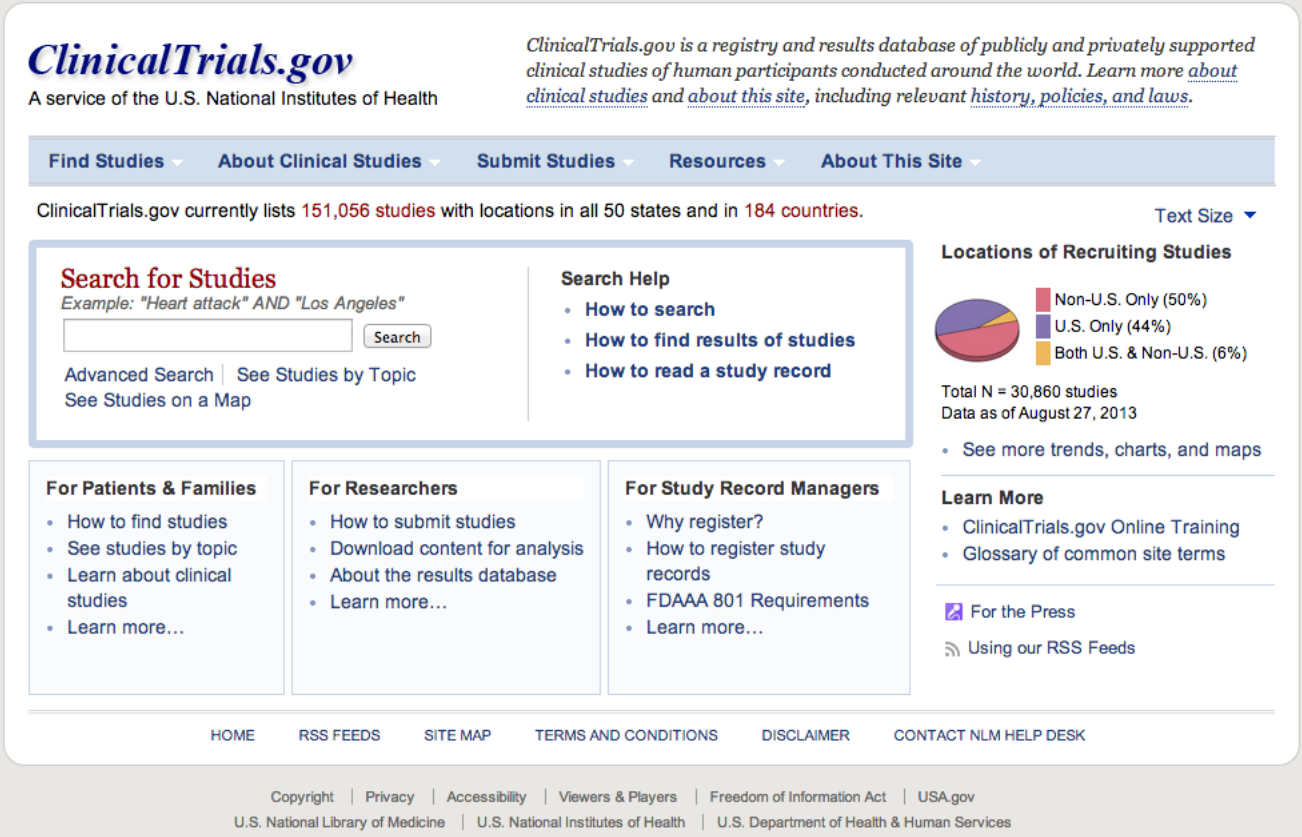
Linked Government Data: datos.gob.cl



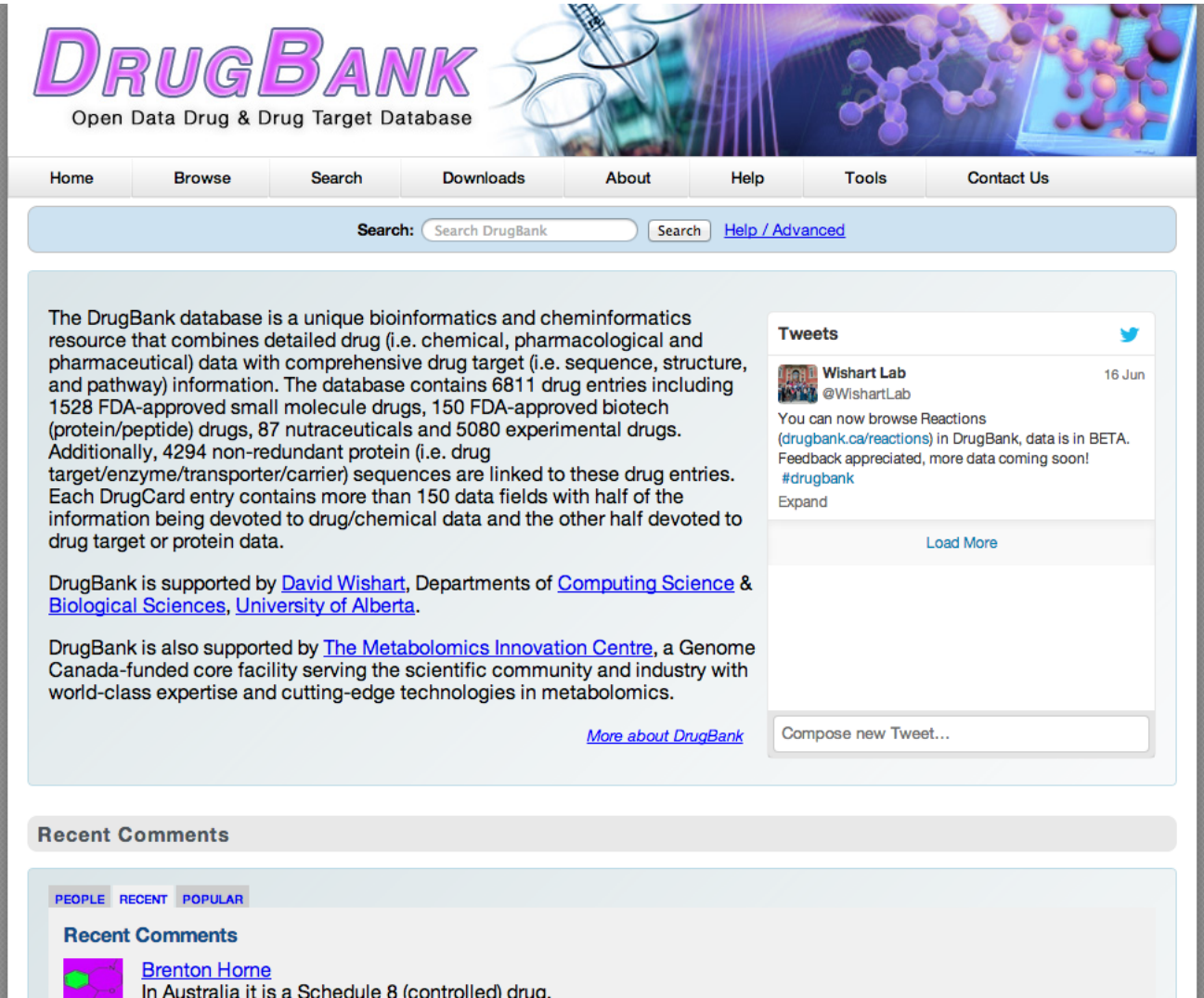
Life Sciences



Life Sciences



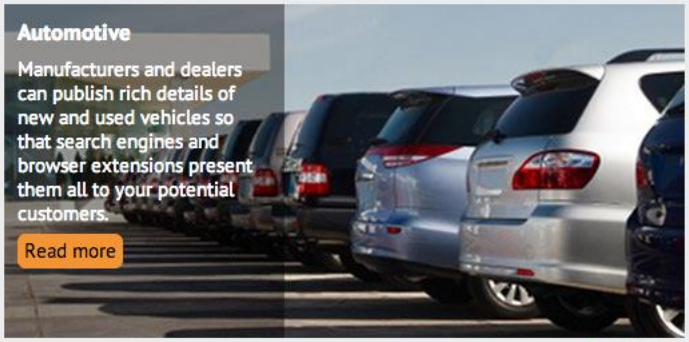
Life Sciences



E-Commerce

GoodRelations
The Web Vocabulary for E-Commerce

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



<<Prev Next>>

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?

Google
Yahoo!
BestBuy
sears.com
kmart.com

... and 10,000 more

See here for additional references.

News from Twitter

Contact

Univ.-Prof. Dr. Martin Hepp

License

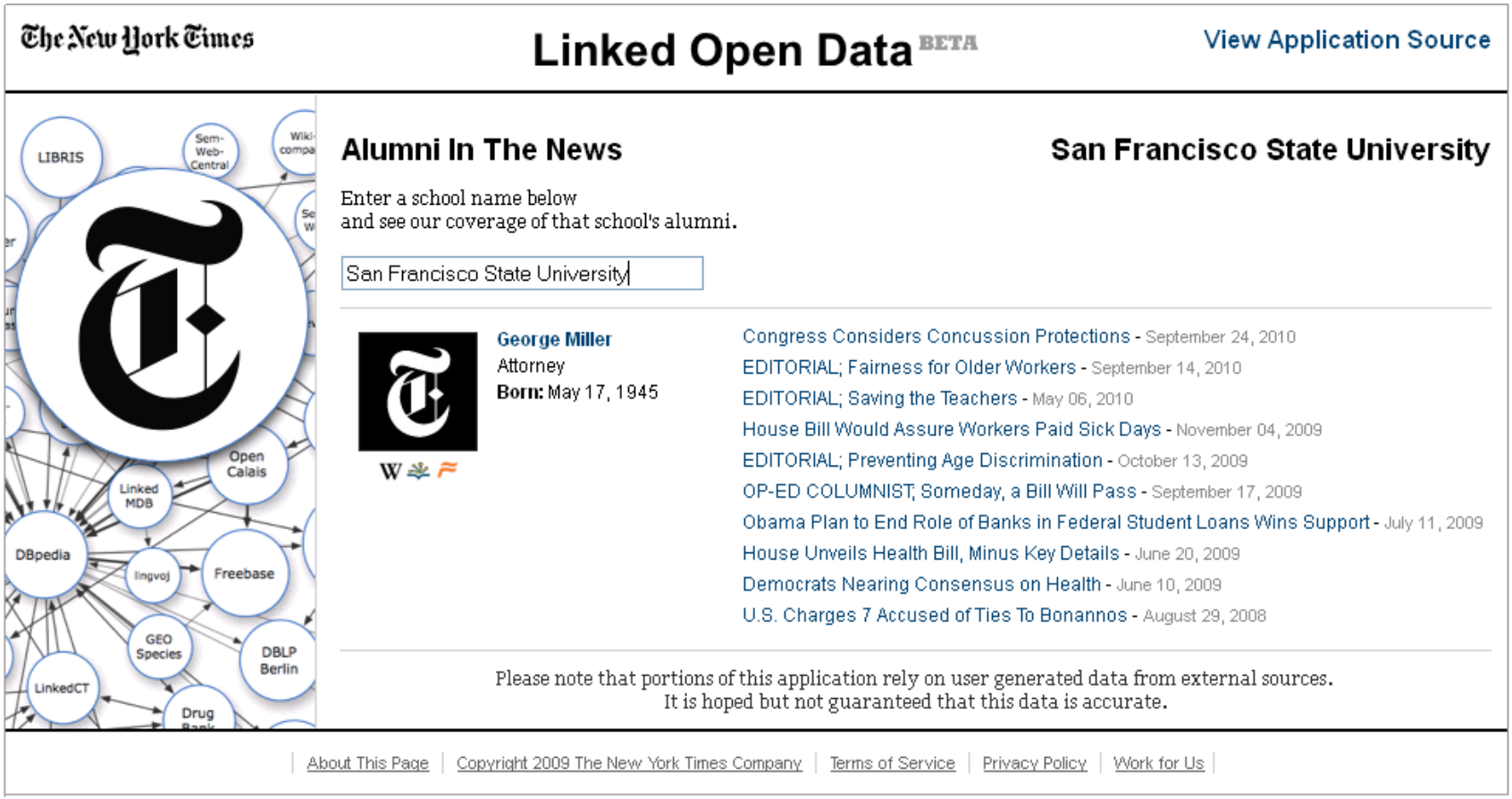
The GoodRelations ontology is creative

Acknowledgments

Many organizations and individuals have supported

Feedback

New York Times Meta-data



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

BBC Music

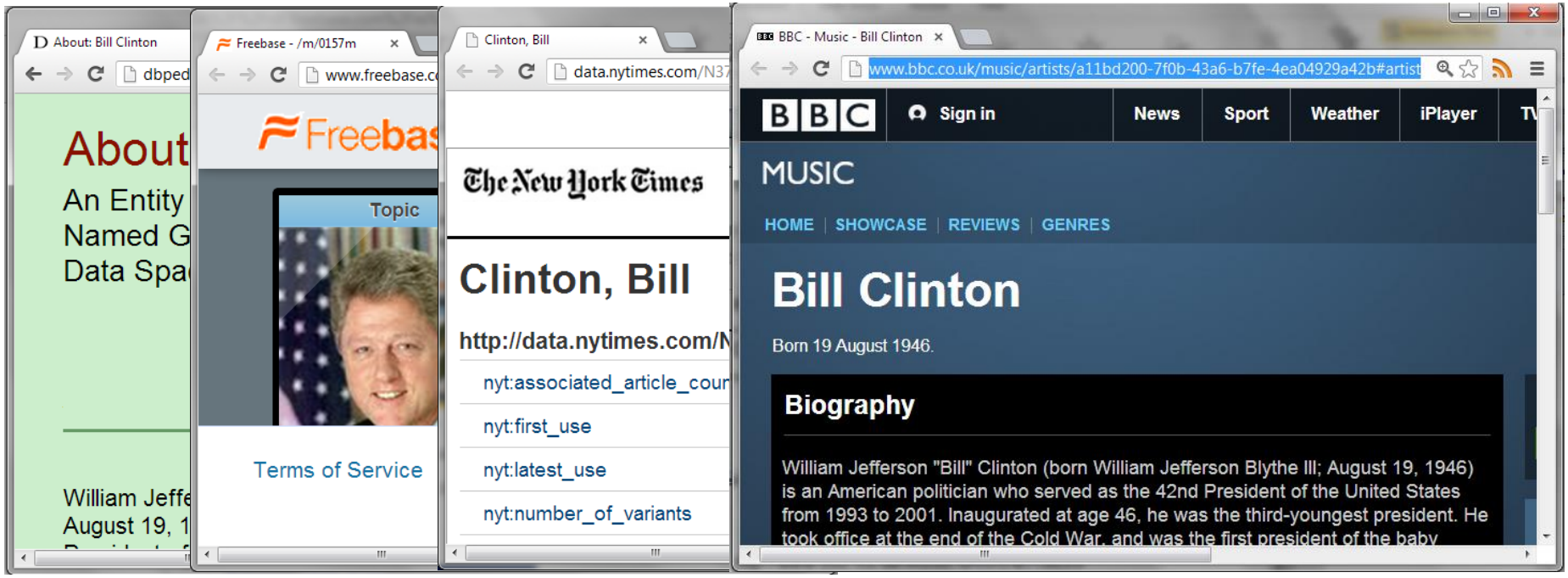

BBC
Sign in
News
Sport
Weather
iPlayer
TV
Radio
More...
London 2012
Search
MUSIC
HOME | SHOWCASE | REVIEWS | GENRES
Search By Artist...
The Beatles
Formed 1957. Disbanded 10 April 1970.
PLAYED MOST ON BBC RADIO 2
BBC TWO
ARENA | BBC TWO
Brian Epstein Finds The Beatles
Brian Epstein, Paul McCartney and John Lennon talk about the early days of The Beatles in the Arena documentary, The Brian Epstein Story (1998)
THE BEATLES 1-4 OF 15
Featured on BBC MUSIC SHOWCASE
RINGO STARR ON THE BEATLES...
RINGO STARR ON PLAYING WITH THE...
THE BEATLES AND BRIAN...
BRIAN EPSTEIN FINDS THE...
Share This Page 15 so far
Share
facebook
twitter
BBC Music Showcase
BBC MUSIC SHOWCASE
Watch and listen to exclusive music clips
Latest Tracks Played On The BBC
Twist & Shout
BBC RADIO 2 | ZOE BALL AND RICHARD BACON GO FOR GOLD 01/08/2012
Tomorrow Never Knows
BBC 6 MUSIC | RADCLIFFE AND MACONIE WEDNESDAY - JOHNNY SHARP
She Loves You
BBC RADIO 2 | KEN BRUCE CLAUDIA WINKLEMAN SITS IN
Sgt Pepper's Lonely Hearts/With A Little Help
BBC RADIO 2 | STEVE WRIGHT IN THE AFTERNOON PATRICK KIELTY SITS IN

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

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
   ?p1 ex:cites ?p2 .
   GRAPH :dblp { ?p1 a :Paper . ?p2 a :Paper }
   ?p1 ex:writtenBy ?a . ?a ex:basedIn wiki:Chile .
   NOT EXISTS {
     ?p1 ex:writtenBy ?b . ?p2 ex:writtenBy ?b .
   }
} GROUP BY ?a ORDER BY DESC(?c)
```

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

Citation indices

	All	Since 2009
Citations	4356	2488
h-index	30	25
i10-index	47	44

Show: 20 1-20 Next >

Title / Author	Cited by	Year
Consistent query answers in inconsistent databases M Arenas, L Bertossi, J Chomicki Proceedings of the eighteenth ACM SIGMOD-SIGACT-SIGART symposium on ...	584	1999
A normal form for XML documents M Arenas, L Libkin ACM Transactions on Database Systems (TODS) 29 (1), 195-232	469	2004

dblp .uni-trier.de Computer Science Bibliography

Universität Trier

Marcelo Arenas

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

Facets and more with CompleteSearch

Ask others: ACM DL/Guide - - CSB - MetaPress - Google - Bing - Yahoo

author:marcelo_arenas:

2013

j30 Marcelo Arenas, Pablo Barceló, Ronald Fagin, Leonid Libkin: Solutions and query rewriting in data exchange. Inf. Comput. 228: 28-61 (2013)

j29 Marcelo Arenas, Jorge Pérez, Juan L. Reutter: Data exchange beyond complete data. J. ACM 60(4): 28 (2013)

Refine by AUTHOR
Marcelo Arenas (101)
Jorge Pérez (25)
Leonid Libkin (22)
Pablo Barceló (15)
[top 4] [top 50] [all 60]

Refine by VENUE
PODS (10)
CoRR (7)
Encyclopedia of Database Systems (6)
Description Logics (4)
[top 4] [top 50] [all 52]

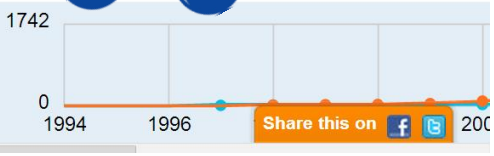
Microsoft Academic Search Beta

Co-authors (93)
Leonid Libkin
Leopoldo Bertossi
Claudio Gutierrez (Claudio Gutiérrez)
Pablo Barcelo (Pablo Barceló)
Jan Chomicki

Citation Graph

Academic > Authors > Marcelo Arenas

celo Arenas
icia Universidad Católica de Chile
Publications: 102 | Citations: 2074
Fields: Databases, Algorithms & Theory,
Collaborated with 93 co-authors from 199



Create account Log in

Article Talk Read

WIKIPEDIA
The Free Encyclopedia

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

Pontifical Catholic University of Chile

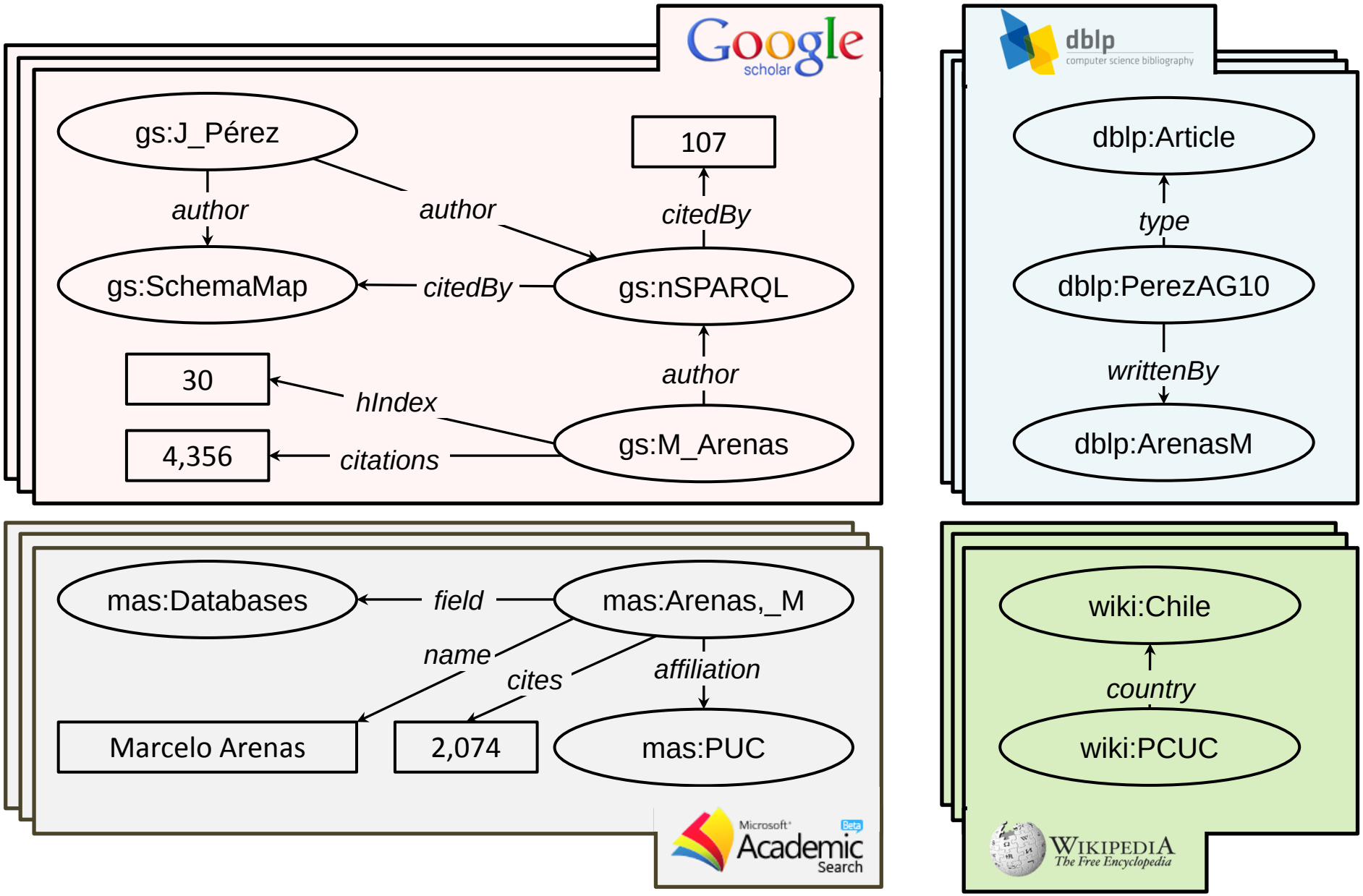
From Wikipedia, the free encyclopedia

Coordinates: 33.4411°S 70.6408°W

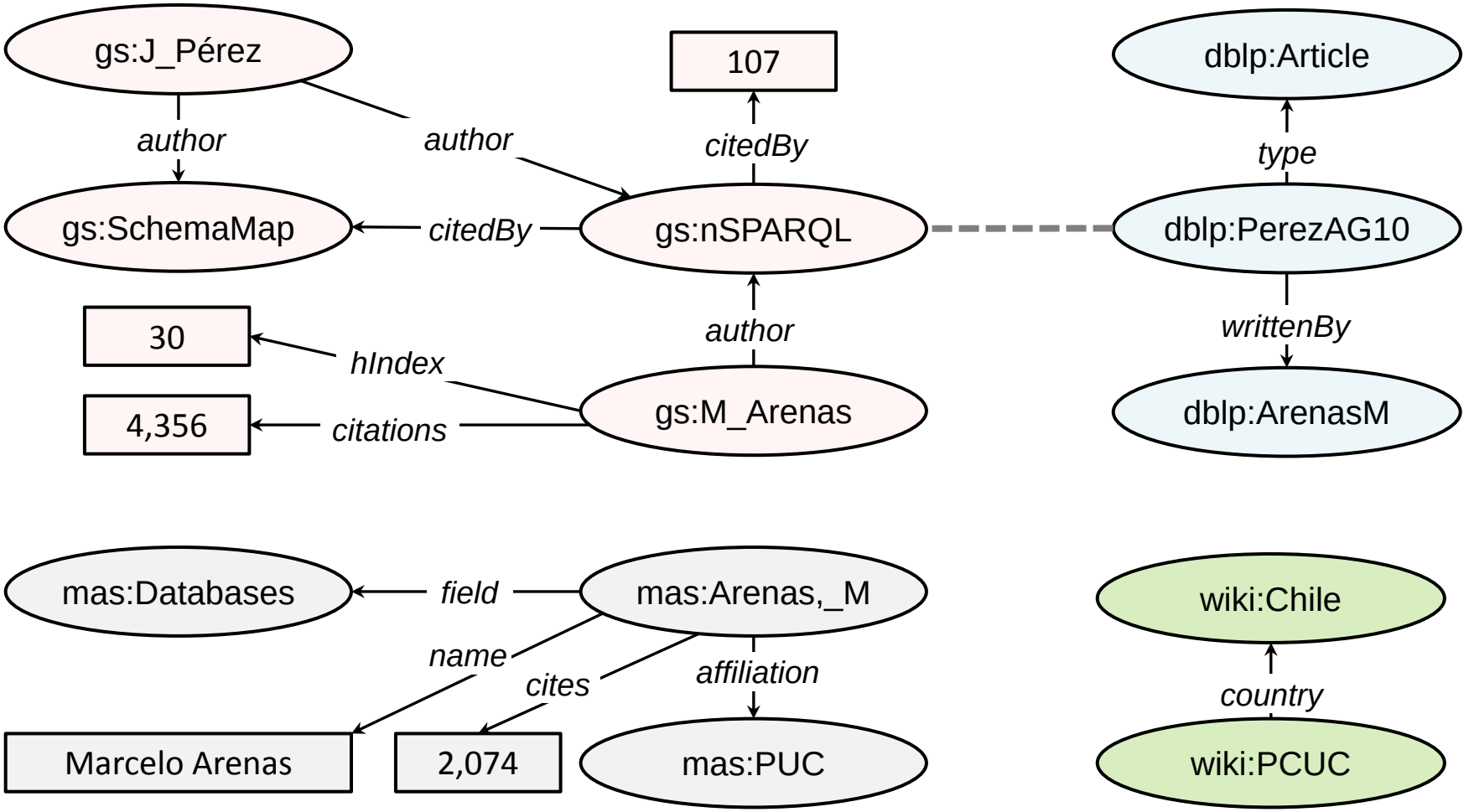
The **Pontifical Catholic University of Chile (UC or PUC)** (Spanish: *Pontificia Universidad Católica de Chile*) is one of the six Catholic Universities existing in the Chilean university system and one of the two Pontifical Universities in the country, along with the Pontifical Catholic University of Valparaíso. It is also one of Chile's oldest

Pontifical Catholic University of Chile
Pontificia Universidad Católica de Chile

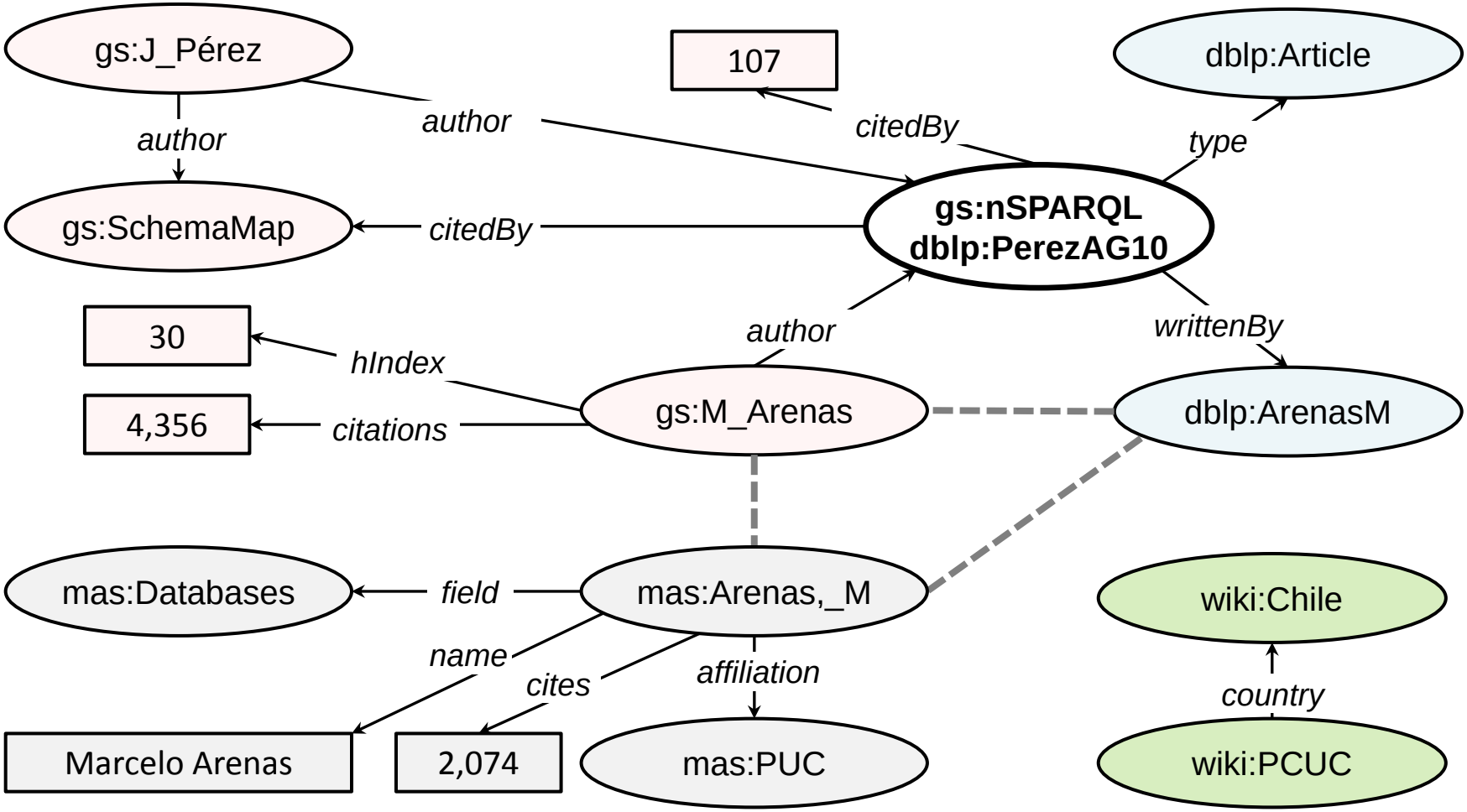
A (Hypothetical) Integration Example



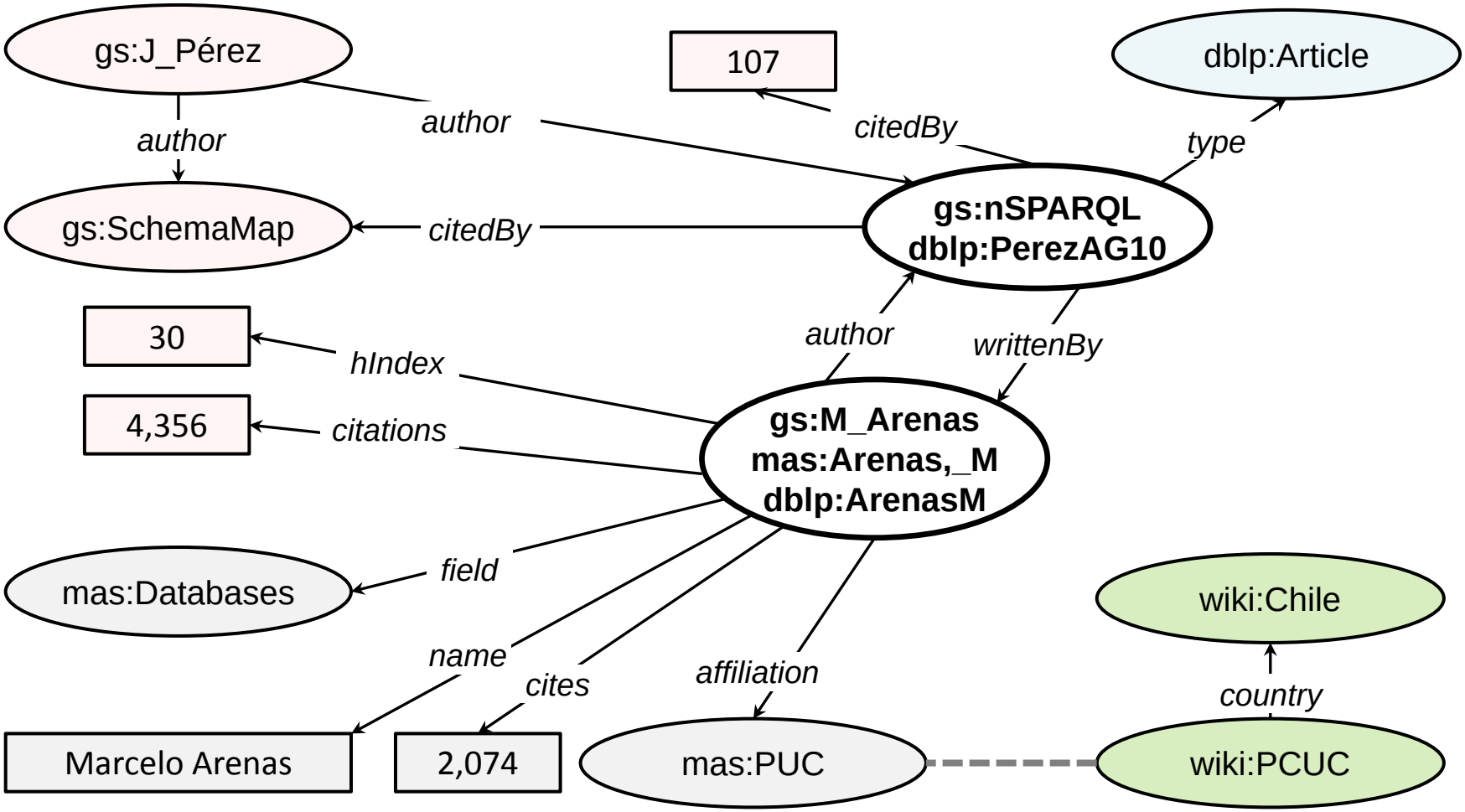
A (Hypothetical) Integration Example



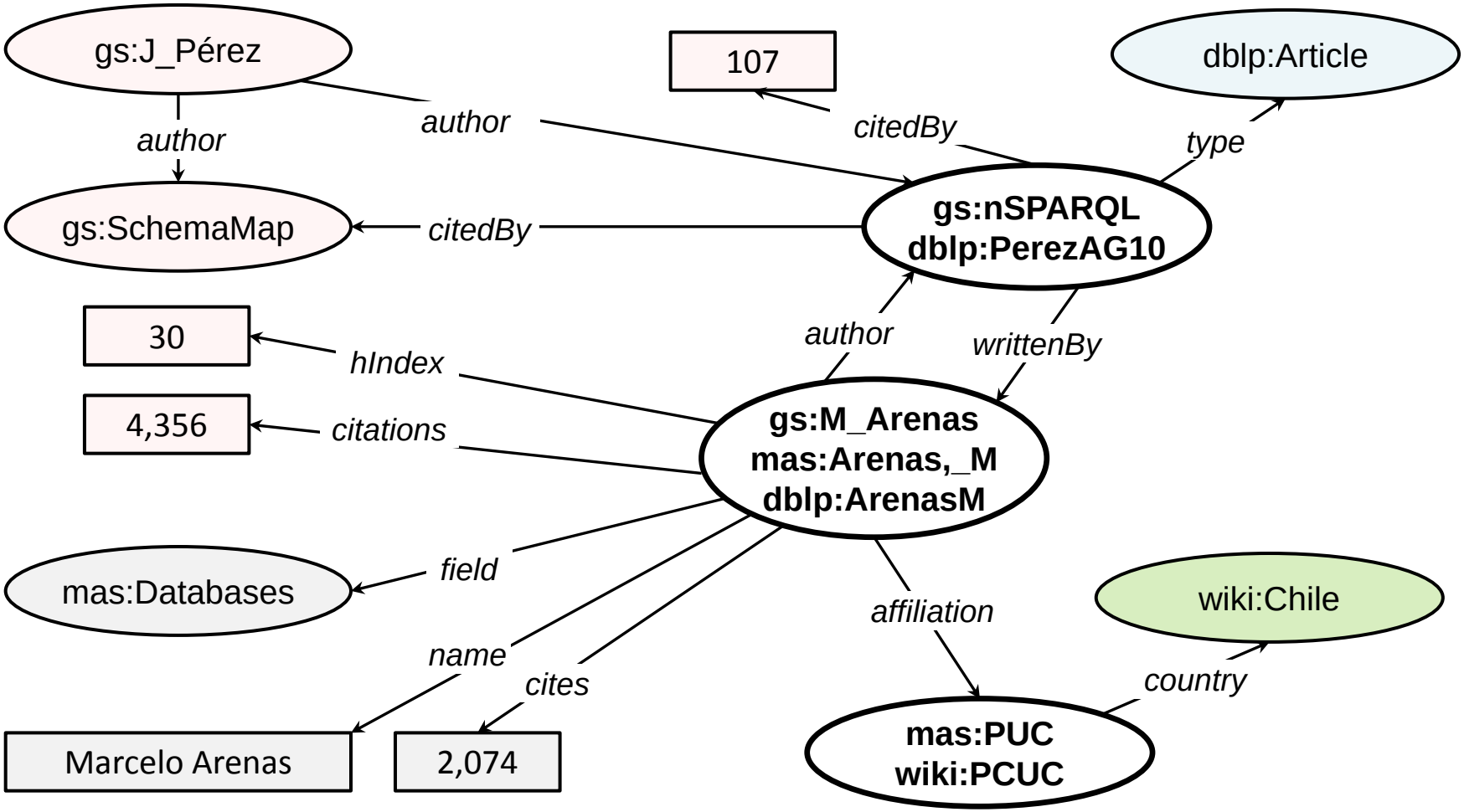
A (Hypothetical) Integration Example



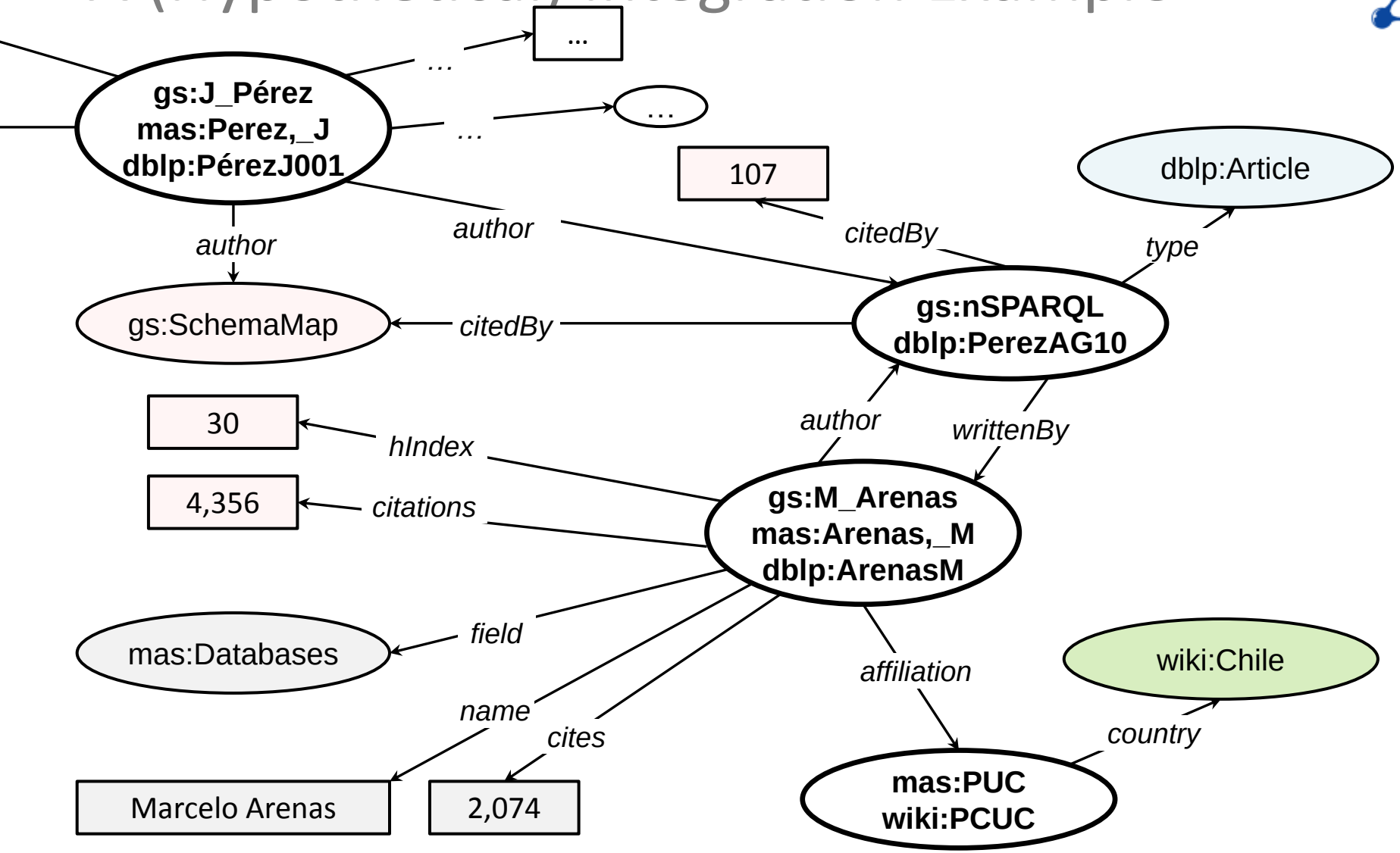
Semantic Web: Tackling Heterogeneity



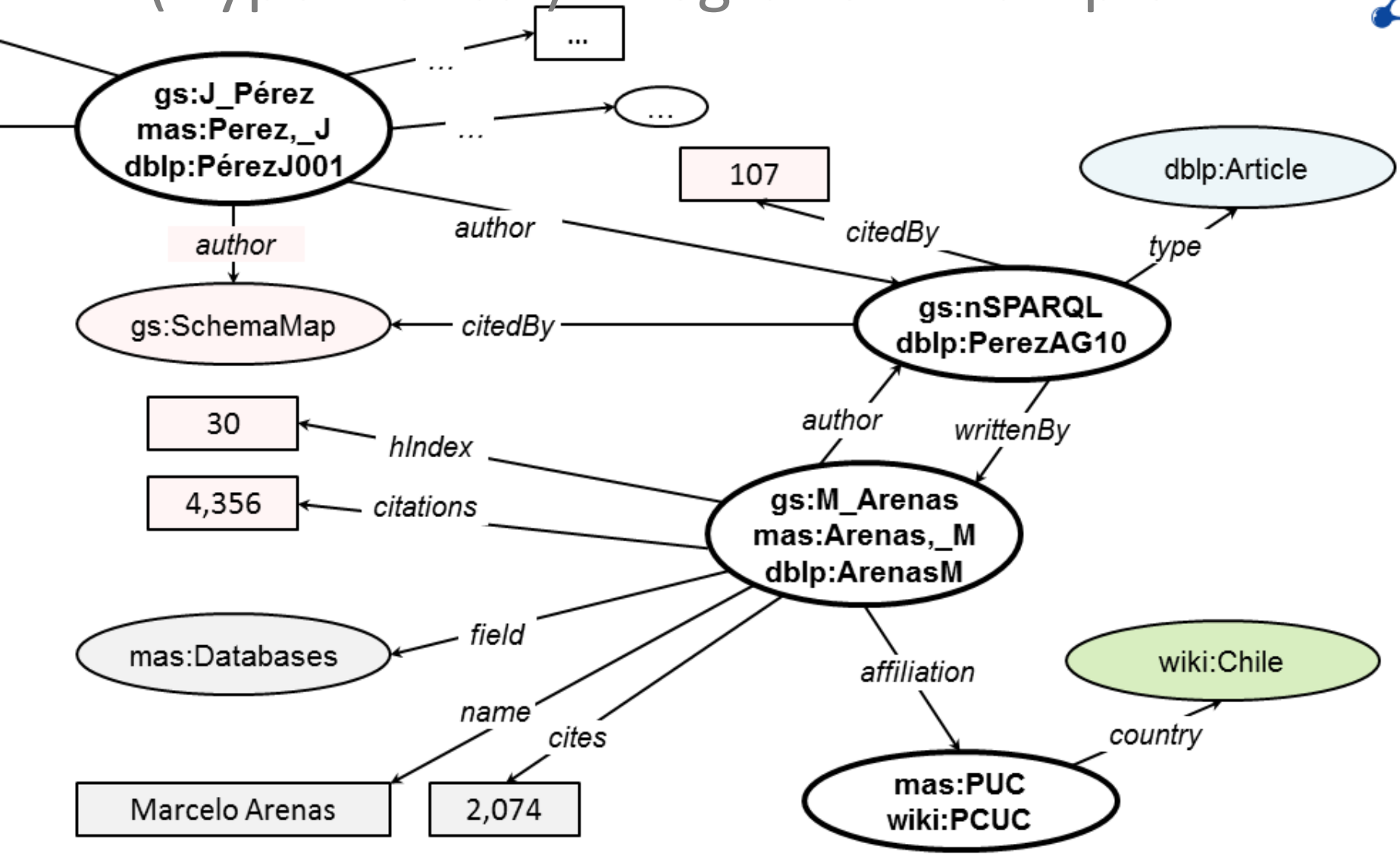
A (Hypothetical) Integration Example



A (Hypothetical) Integration Example

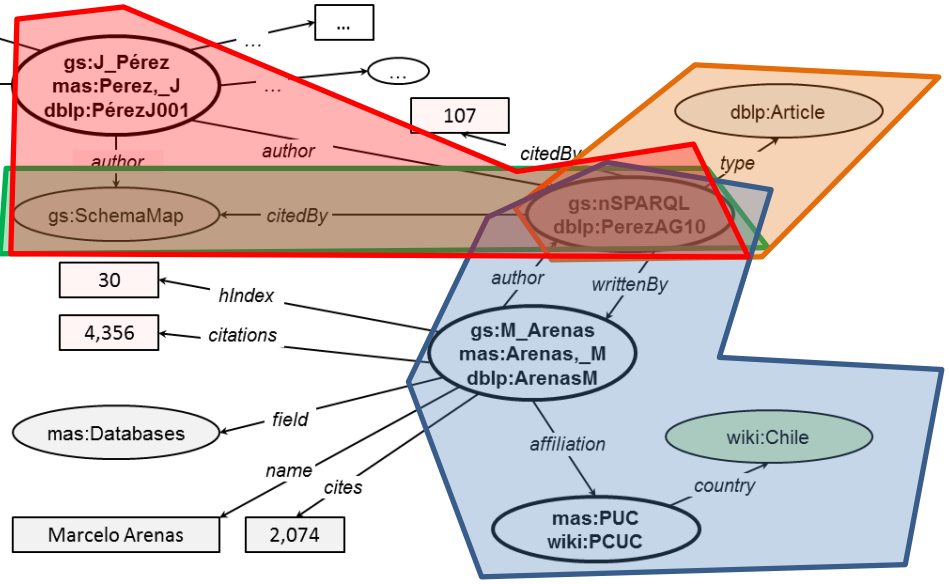


A (Hypothetical) Integration Example



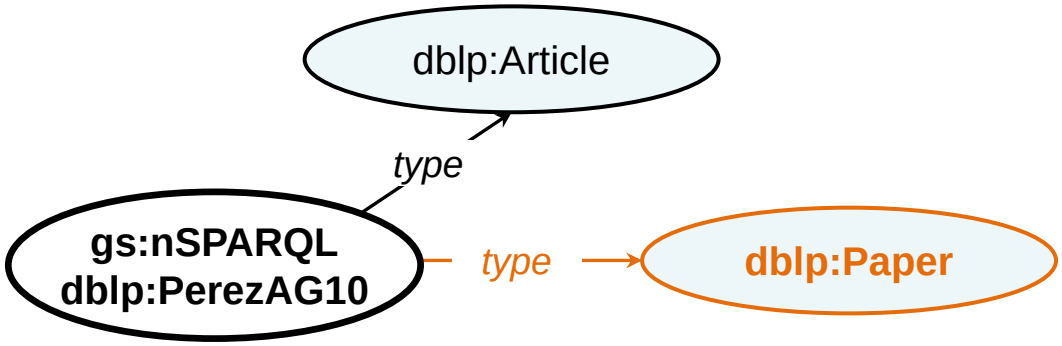
A (Hypothetical) Integration Example

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
    ?p1 ex:citedBy ?p2 .
    GRAPH :dblp { ?p1 a :Paper . ?p2 a :Paper }
    ?p1 ex:writtenBy ?a . ?a ex:basedIn wiki:Chile .
    NOT EXISTS {
      ?p1 ex:writtenBy ?b . ?p2 ex:writtenBy ?b .
    }
} GROUP BY ?a ORDER BY DESC(?c)
```



A (Hypothetical) Integration Example

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
    ?p1 ex:citedBy ?p2 .
    GRAPH :dblp { ?p1 a :Paper . ?p2 a :Paper }
    ?p1 ex:writtenBy ?a . ?a ex:basedIn wiki:Chile .
    NOT EXISTS {
      ?p1 ex:writtenBy ?b . ?p2 ex:writtenBy ?b .
    }
} GROUP BY ?a ORDER BY DESC(?c)
```

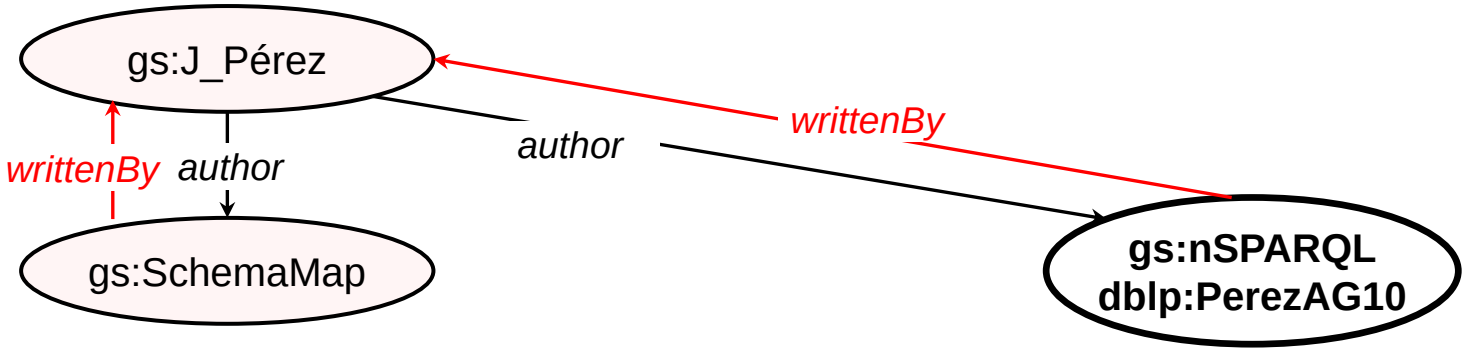


Article ⊑ Paper

```
dblp:Article rdfs:subClassOf dblp:Paper .
```

A (Hypothetical) Integration Example

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
    ?p1 ex:citedBy ?p2 .
    GRAPH :dblp { ?p1 a :Paper . ?p2 a :Paper }
    ?p1 ex:writtenBy ?a . ?a ex:basedIn wiki:Chile .
    NOT EXISTS {
      ?p1 ex:writtenBy ?b . ?p2 ex:writtenBy ?b .
    }
} GROUP BY ?a ORDER BY DESC(?c)
```

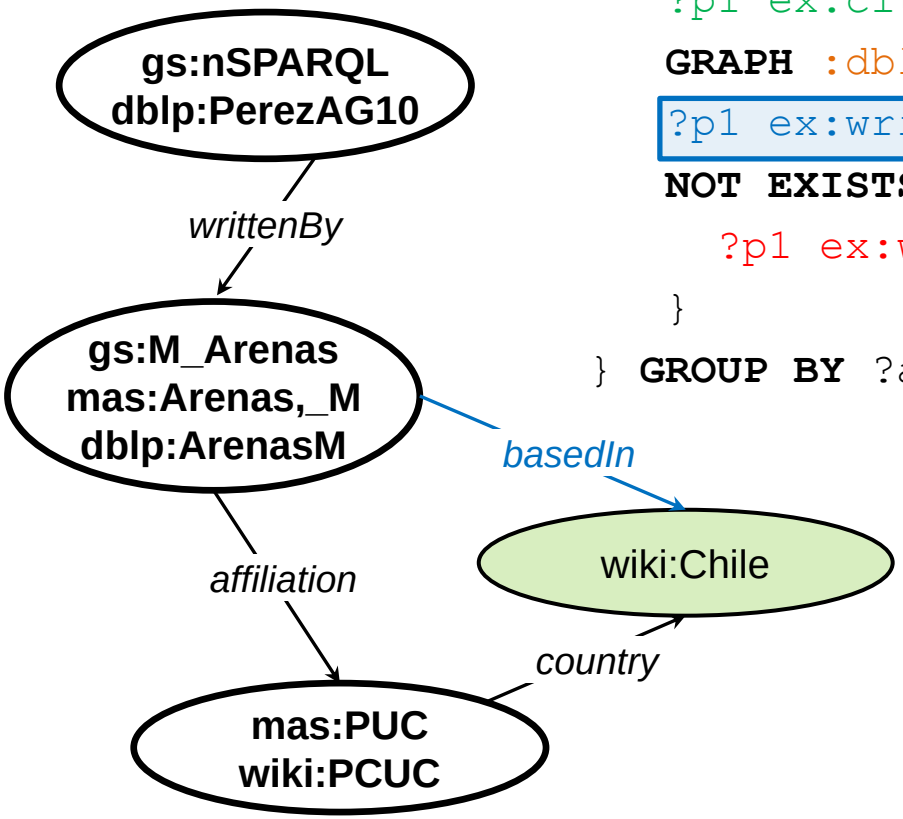


author ≡ writtenBy⁻

```
ex:author owl:inverseOf ex:writtenBy .
```

A (Hypothetical) Integration Example

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
    ?p1 ex:citedBy ?p2 .
    GRAPH :dblp { ?p1 a :Paper . ?p2 a :Paper }
    ?p1 ex:writtenBy ?a . ?a ex:basedIn wiki:Chile .
    NOT EXISTS {
      ?p1 ex:writtenBy ?b . ?p2 ex:writtenBy ?b .
    }
} GROUP BY ?a ORDER BY DESC(?c)
```

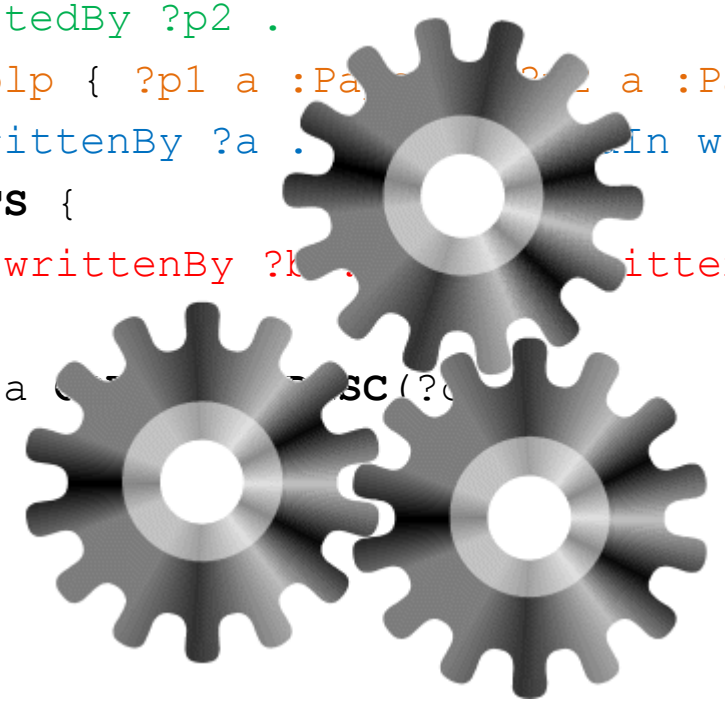


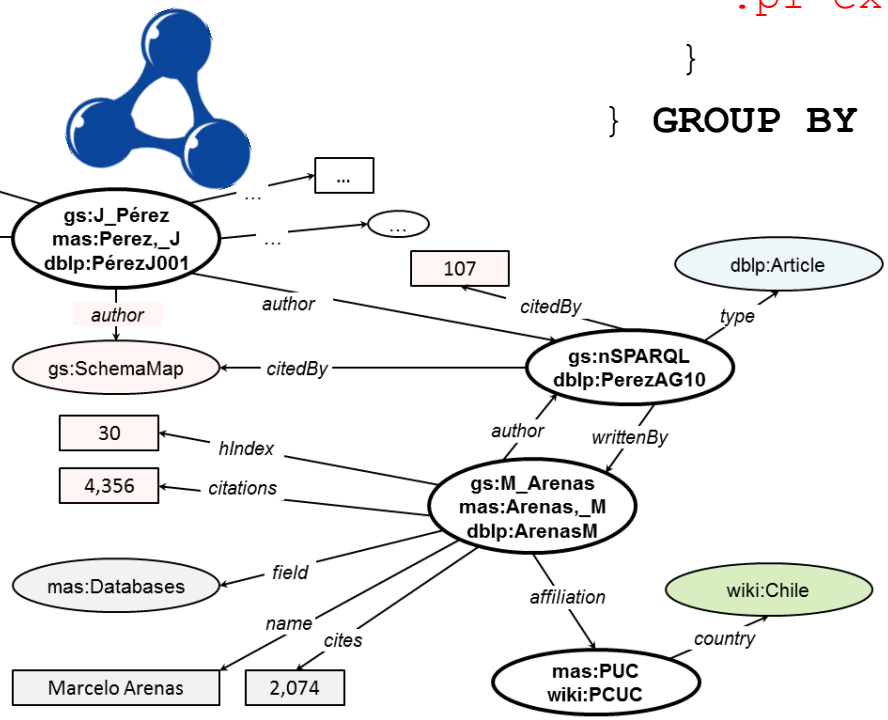
affiliation ∘ country ⊑ basedIn

```
ex:basedIn owl:propertyChainAxiom( ex:affilition ex:country)
```

A (Hypothetical) Integration Example

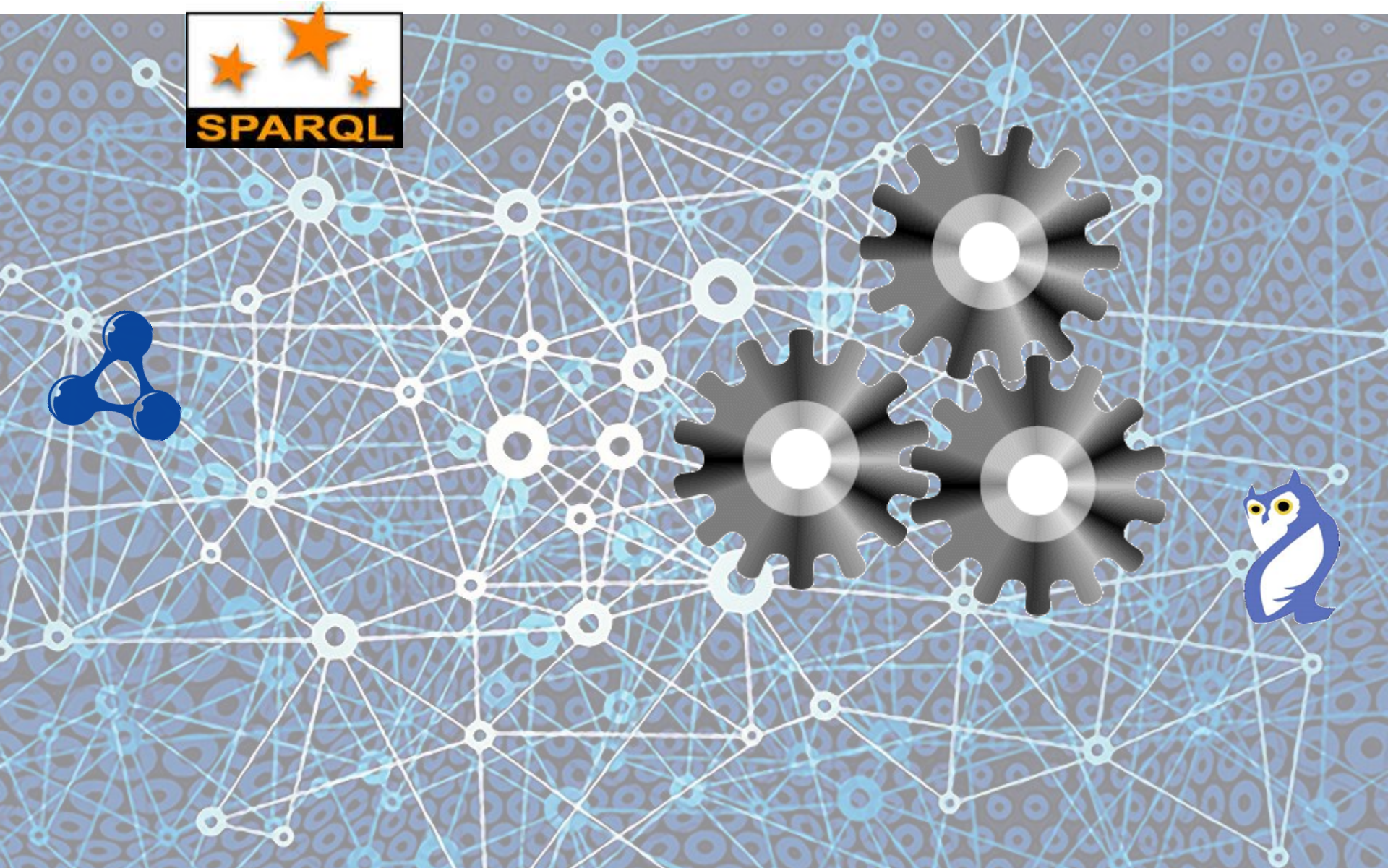

```
SELECT ?a (COUNT(DISTINCT ?p2) AS ?c)
FROM NAMED …
WHERE {
   ?p1 ex:citedBy ?p2 .
   GRAPH :dblp { ?p1 a :Pa…  …2 a :Paper }
   ?p1 ex:writtenBy ?a .  …In wiki:Chile .
   NOT EXISTS {
     ?p1 ex:writtenBy ?k .  …ittenBy ?b .
   }
} GROUP BY ?a …SC(?c…
```





Article ⊑ Paper
author ≡ writtenBy⁻
affiliation ∘ country ⊑ basedIn

Not clear yet how to do reasoning on the Web!



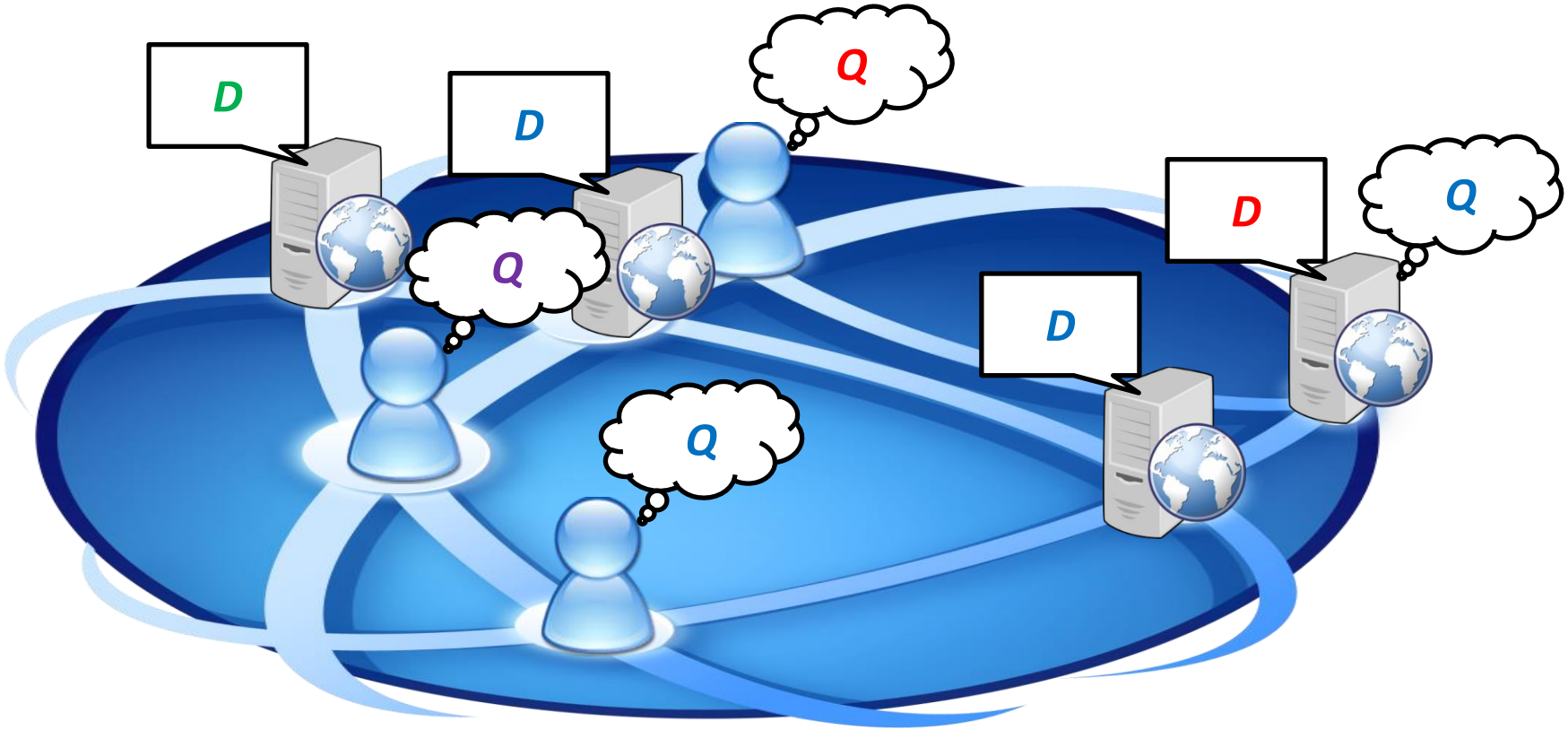
Side Note: Fact or Fiction?

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?)

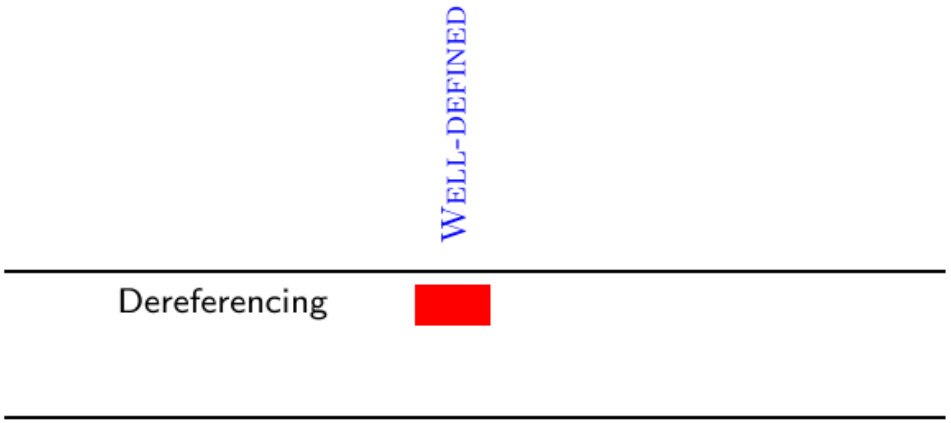
Q = “http://dbpedia.org/resource/Columbia”

Q(*D*) =

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 *Q* and *D*, clients and servers agree on what *Q*(*D*) should be.



Dereferencing (what's wrong with it?)

- **Very coarse:**
  - 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

```
User-agent: *
Crawl-delay: 10
```

GRANULAR: The language for *Q* allows the client to be specific so as to avoid wasting bandwidth

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 into chunks

Dumps (what are they?)

DBpedia Blog | Get Involved | Get Help

DBpedia 3.9 Downloads

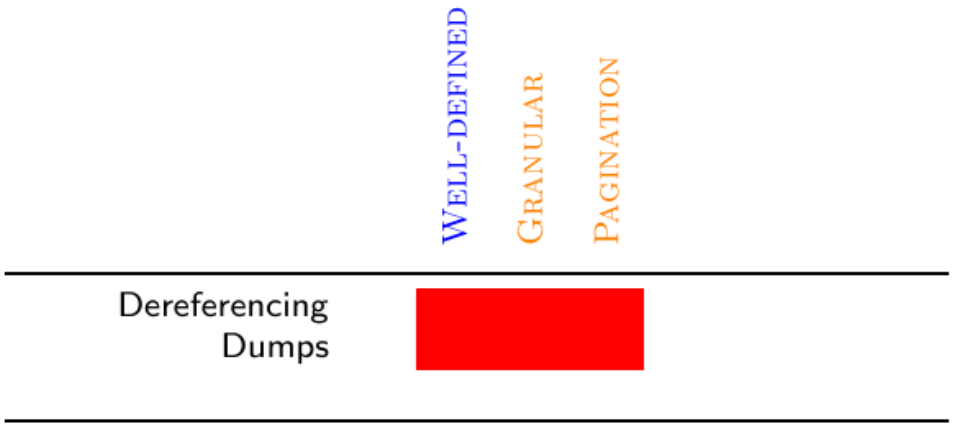
About / News
Applications
Use Cases
Datasets
Online Access
DBpedia Live
Downloads
Interlinking

This pages provides downloads of the DBpedia datasets. The DB
are licensed under the terms of the Creative Commons Attributio
and the GNU Free Documentation License. OPEN DATA The dow
as N-Triples and N-Quads, where the N-Quads version contains
information for each statement. All files are bzip2 [1] packed.
In addition to the RDF version of the data, we also provide a tabu
of the core DBpedia data sets as CSV and JSON files. See DBp

Older Versions: DBpedia 3.8, DBpedia 3.7, DBpedia 3.6, DBped
DBpedia 3.4, DBpedia 3.3, DBpedia 3.2, DBpedia 3.1, DBpedia
DBpedia 2.0

Dumps (what's wrong with them?)

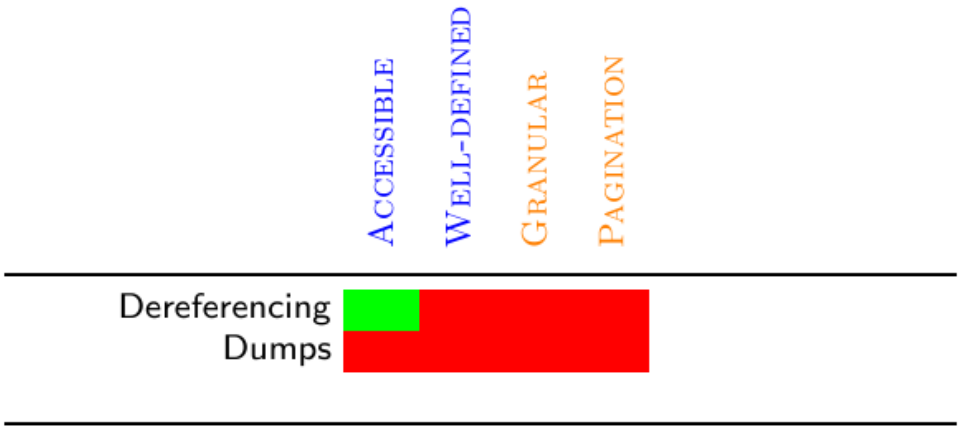
- 15× compression for RDF achievable
- But same weaknesses as for deref. still apply



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 =

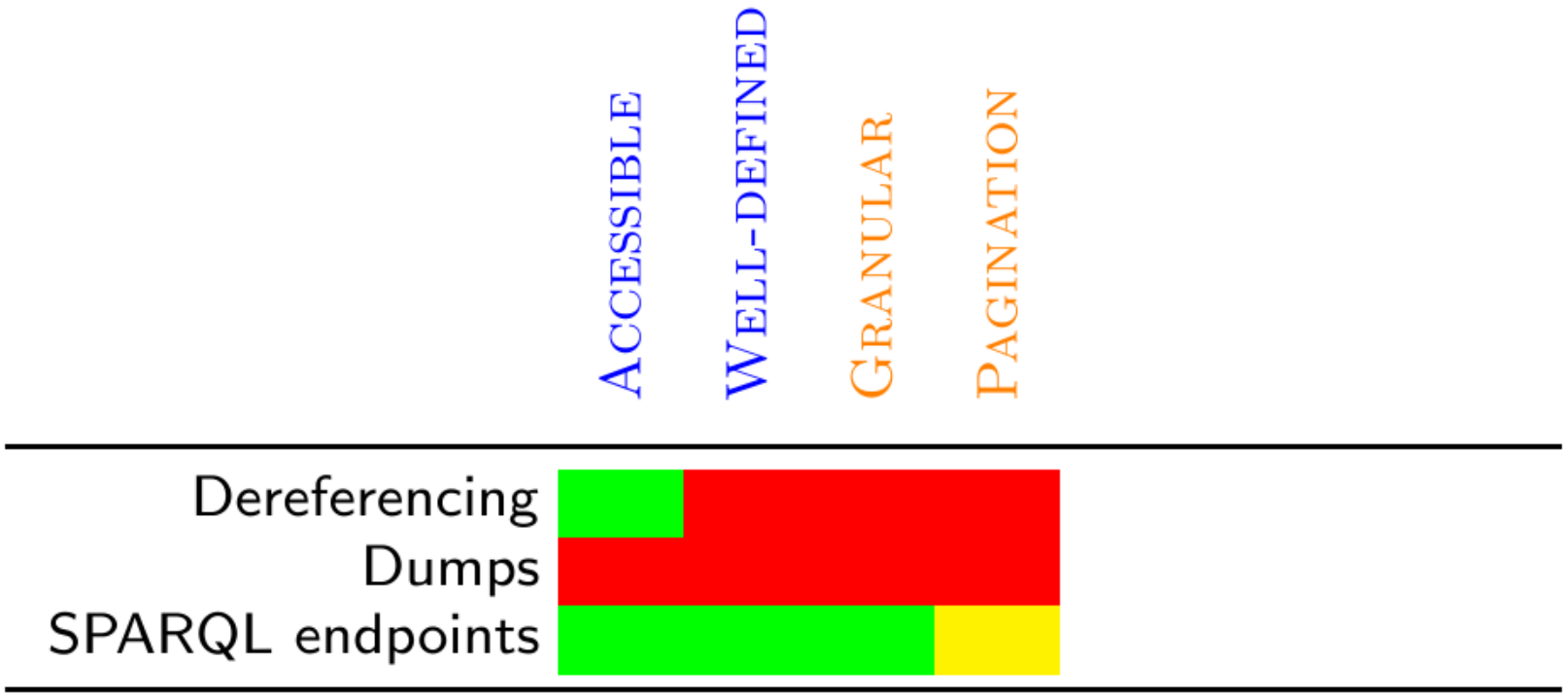
```
PREFIX dbo: <http://dbpedia.org/ontology/>
...

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")
}
```

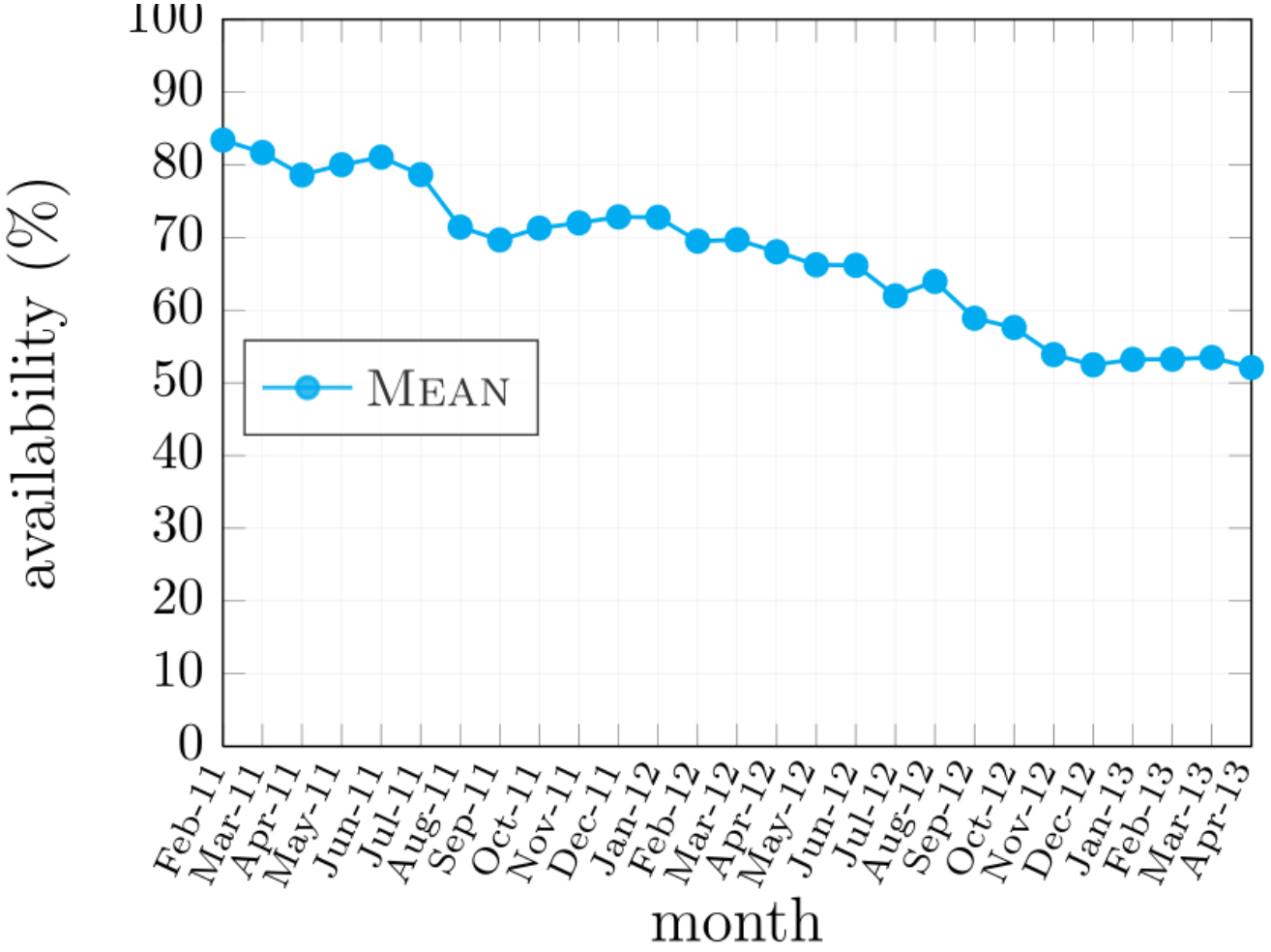
Q(*D*) =

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

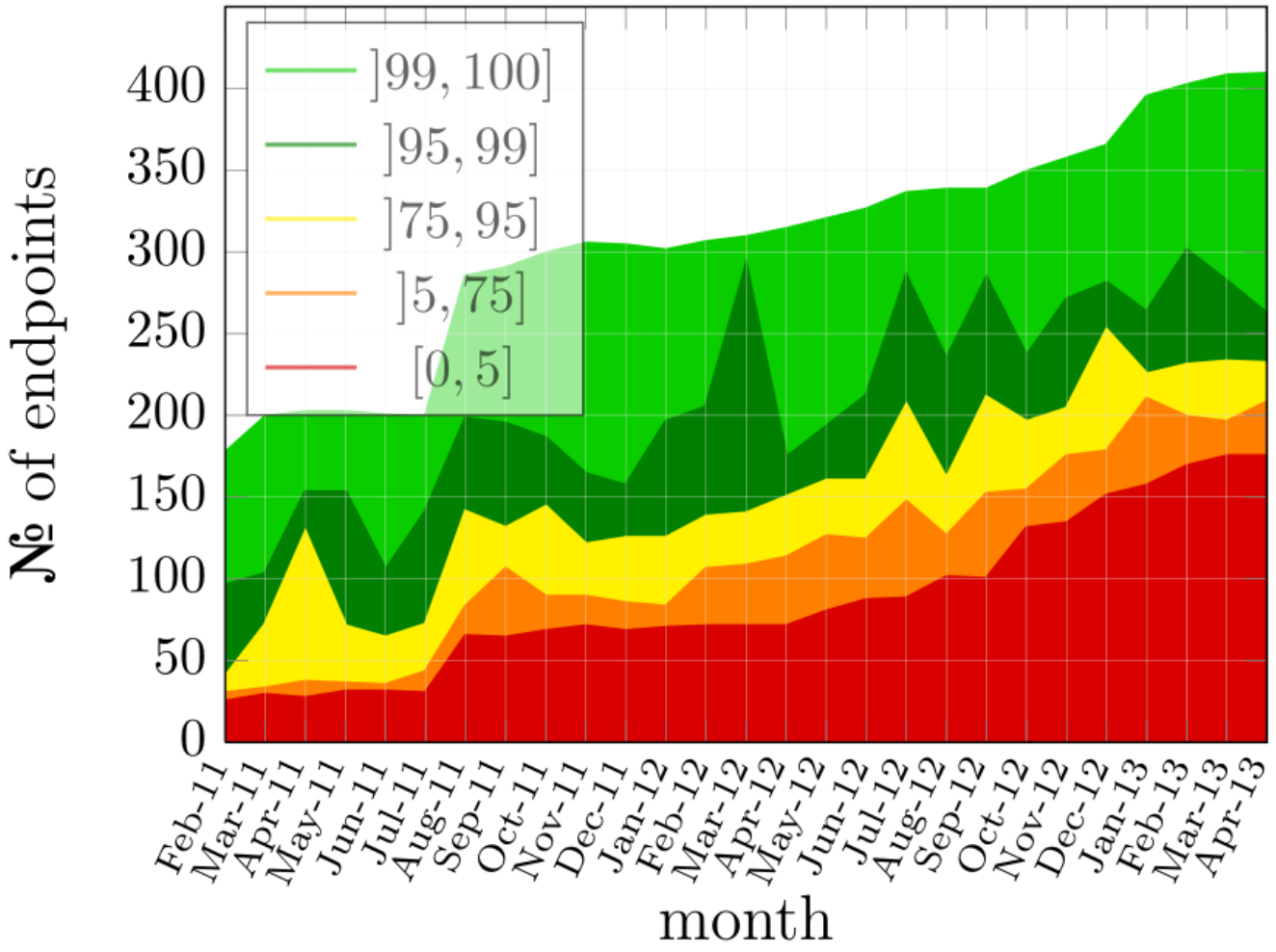
SPARQL (to the rescue?)



SPARQL (what's wrong with it?)

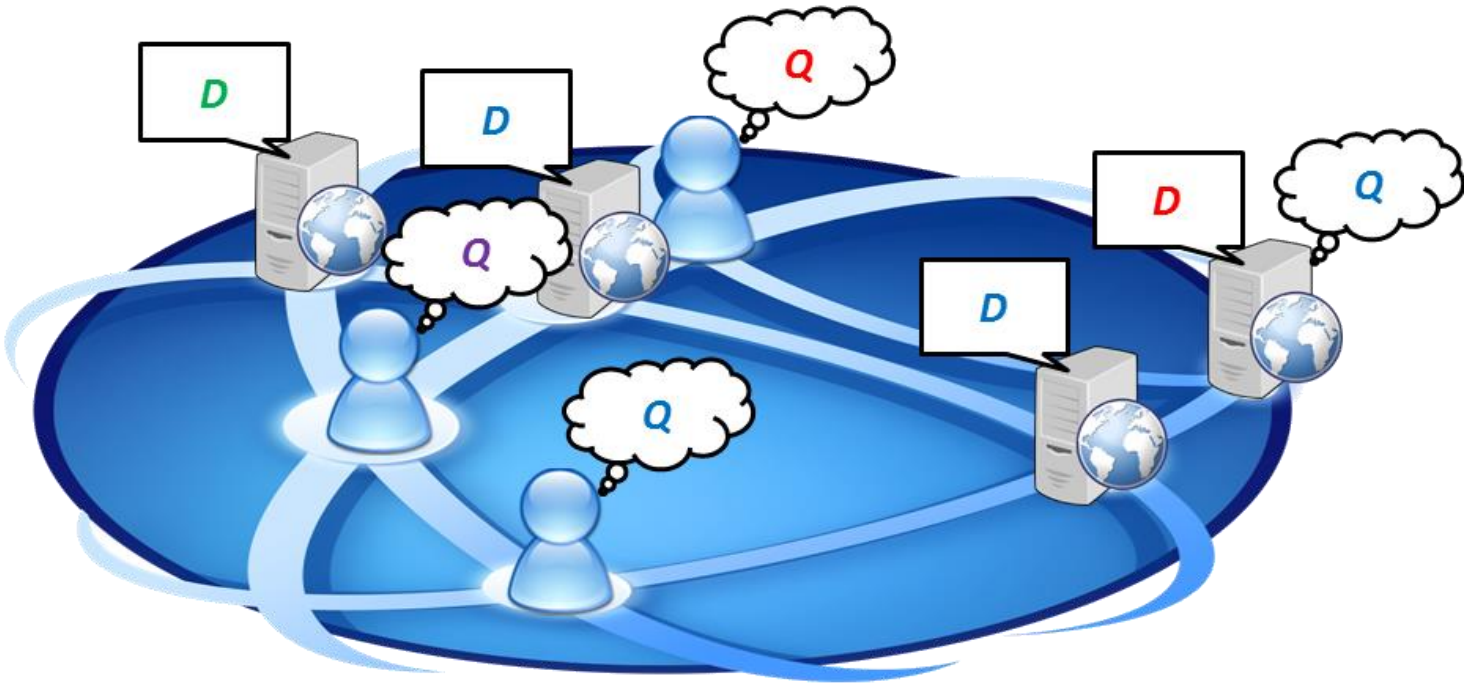


SPARQL (what's wrong with it?)



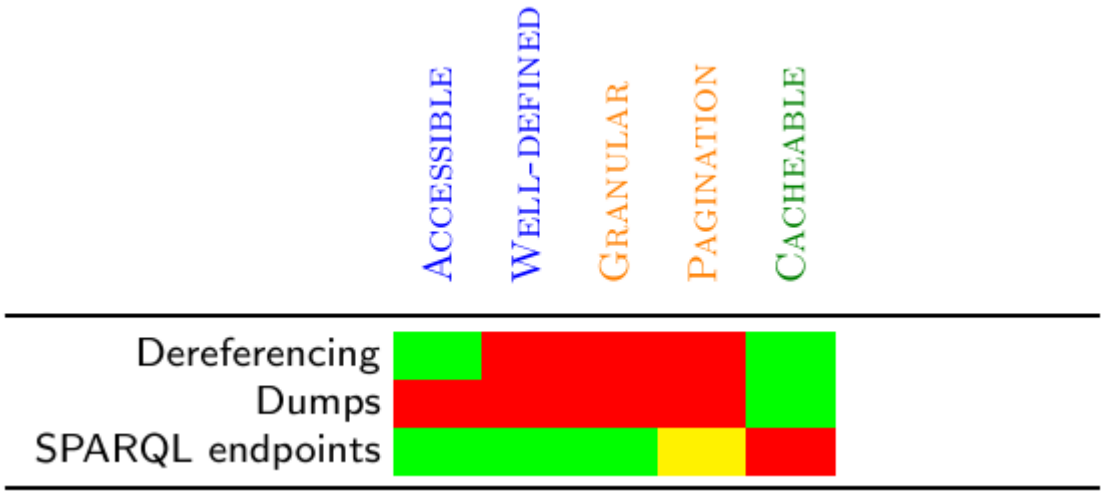
SPARQL (what's wrong with it?)

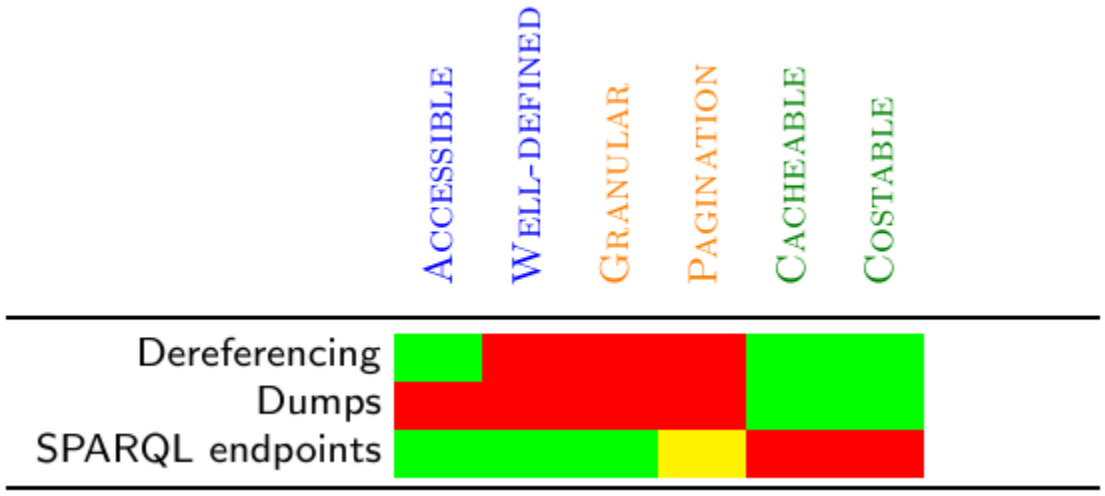
- ~~Simple~~ Protocol and RDF Query Language
- SPARQL evaluation: PSPACE-complete



SPARQL (what's wrong with it?)

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.

SPARQL (what's wrong with it?)

№ of Results	№ 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
```

SPARQL (what's wrong with it?)

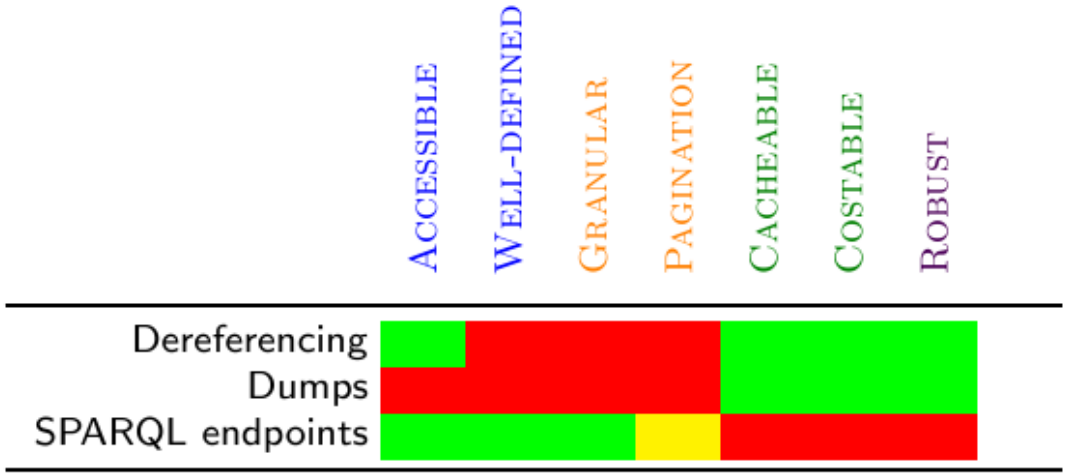
```
Virtuoso 42000 Error The estimated execution time 0
(sec) exceeds the limit of 3000 (sec).

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"))
}
```

SPARQL (what's wrong with it?)

- Simple Protocol And RDF Query Language
- **Protocol always expects a perfect answer**
  - **No support for partial results, timeouts, exception handling, pagination …**

ROBUST: Access method can gracefully handle error cases and provide Quality of Service



SPARQL (what's wrong with it?)

- ***D* is a black-box for the user**

SPARQL (what's wrong with it?)

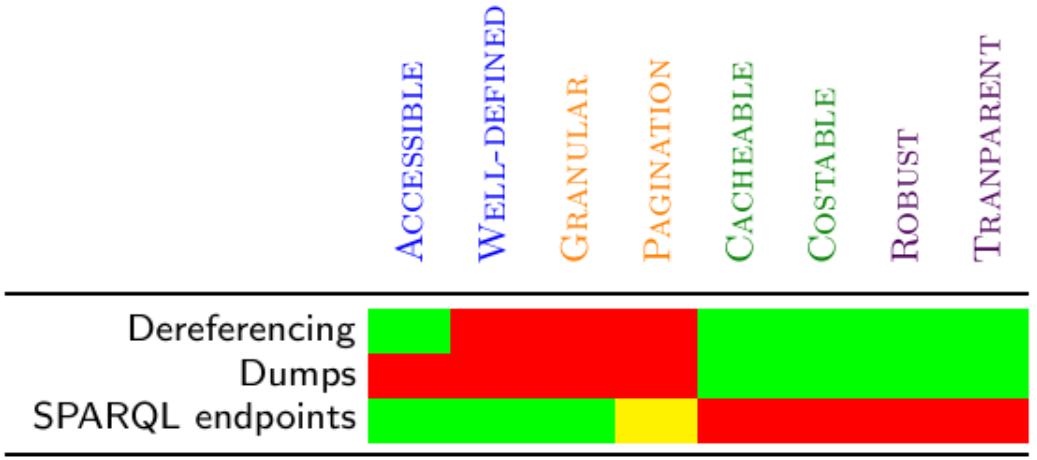
Q =

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

$Q(D)$ =

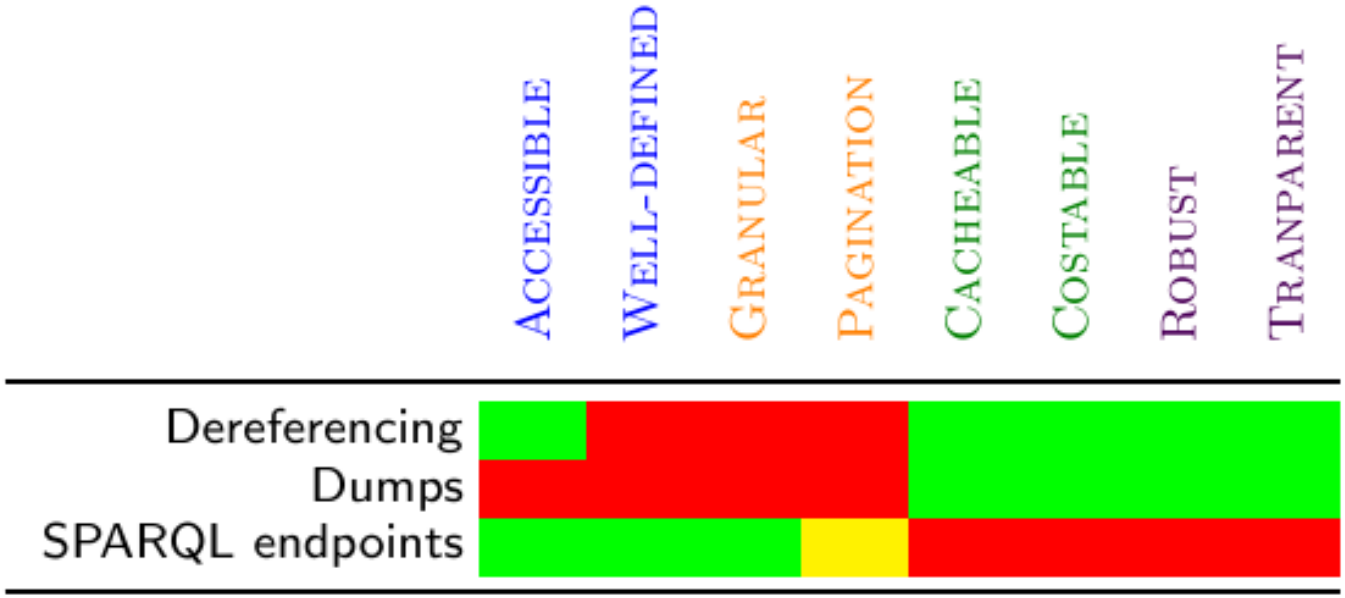
count
7182

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

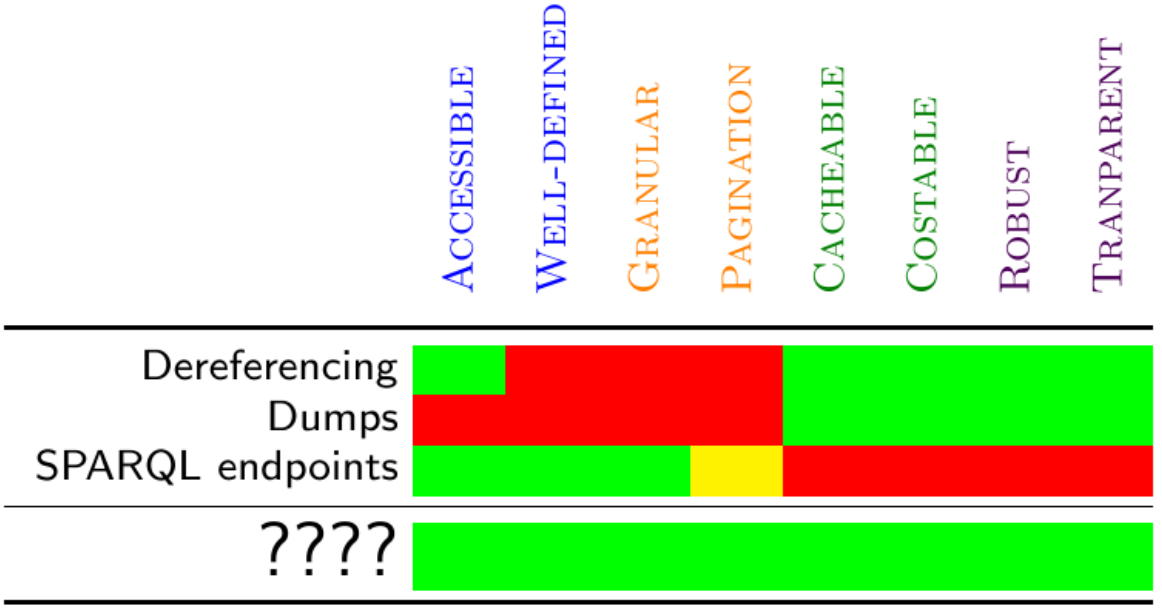


Data Access: Open Problem

- Data access methods crucial
  - Access = Protocol + Query Language
- But current ones don't work

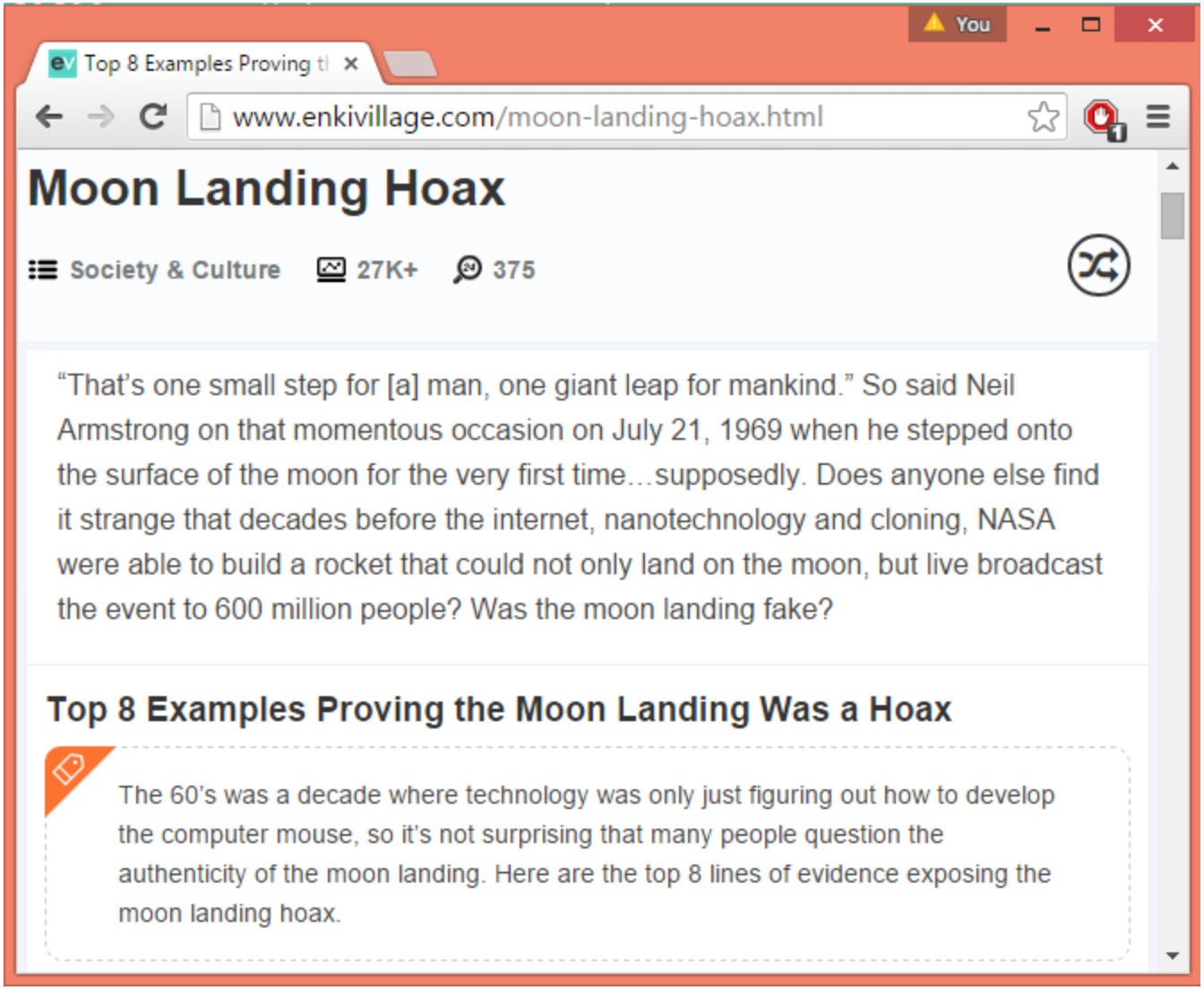
  → *need something else (or multiple things?)*

  – *Open question:*

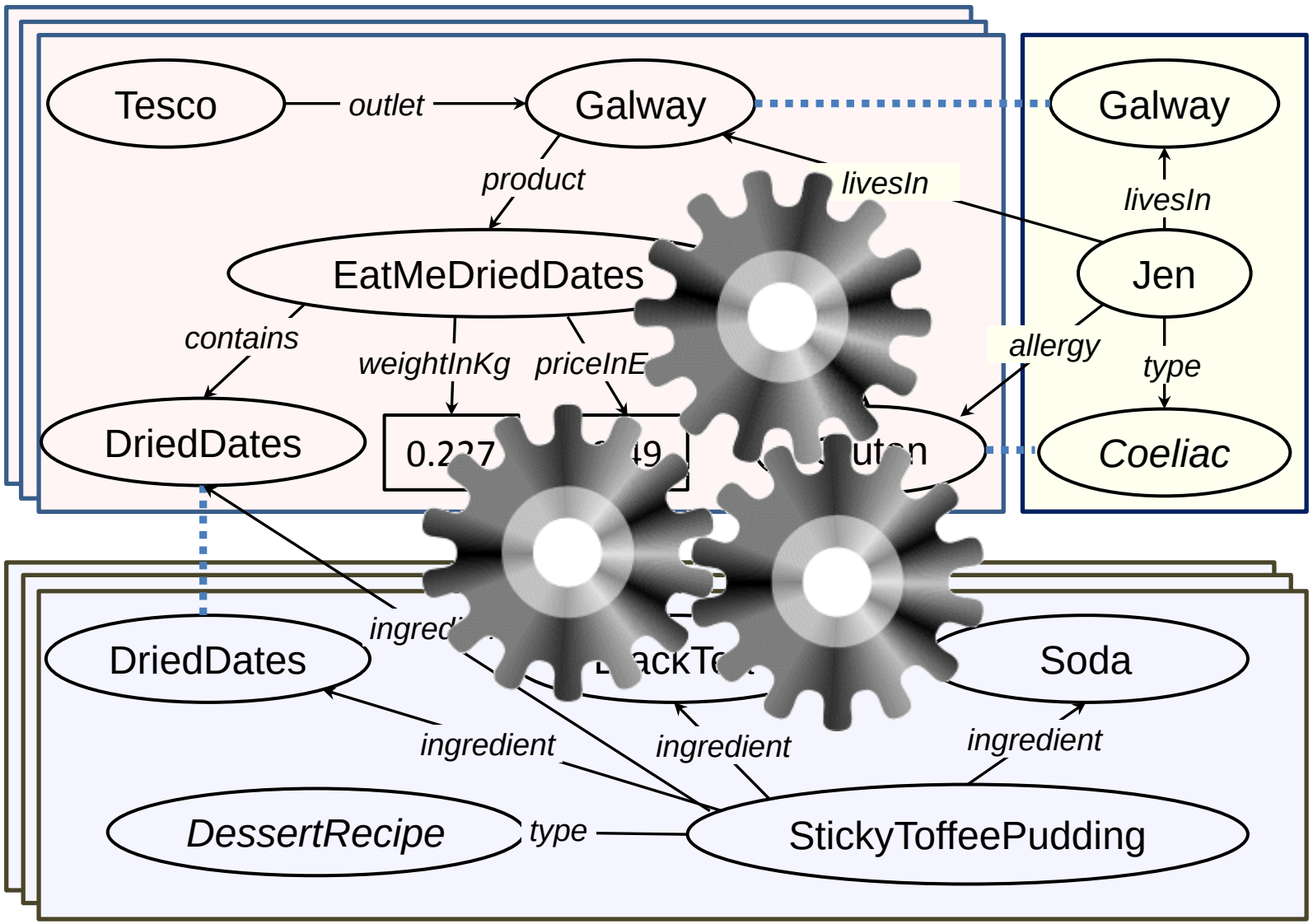


BIG CATCH NUMBER 3: DATA QUALITY

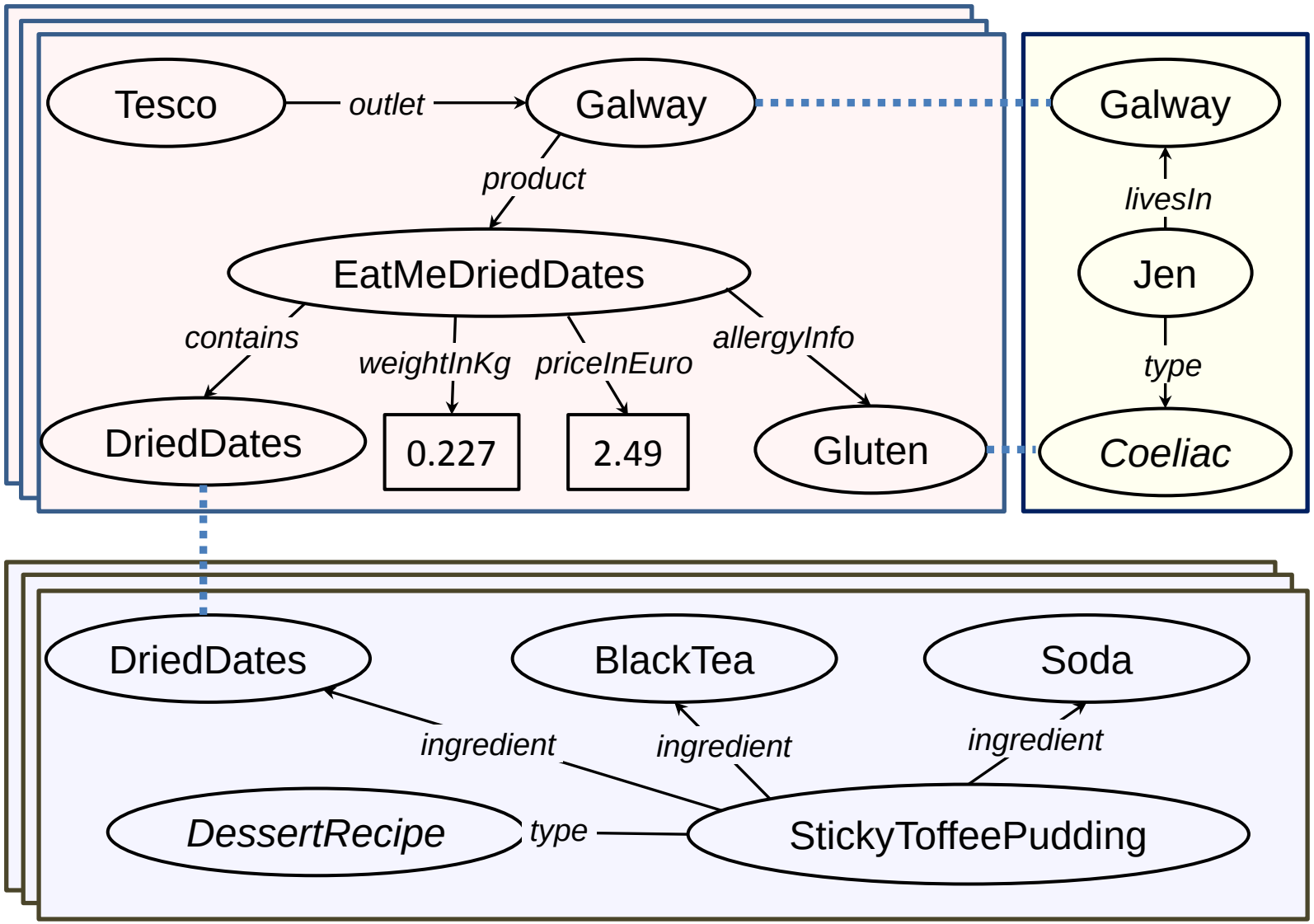
Can’t trust everything you read on the Web



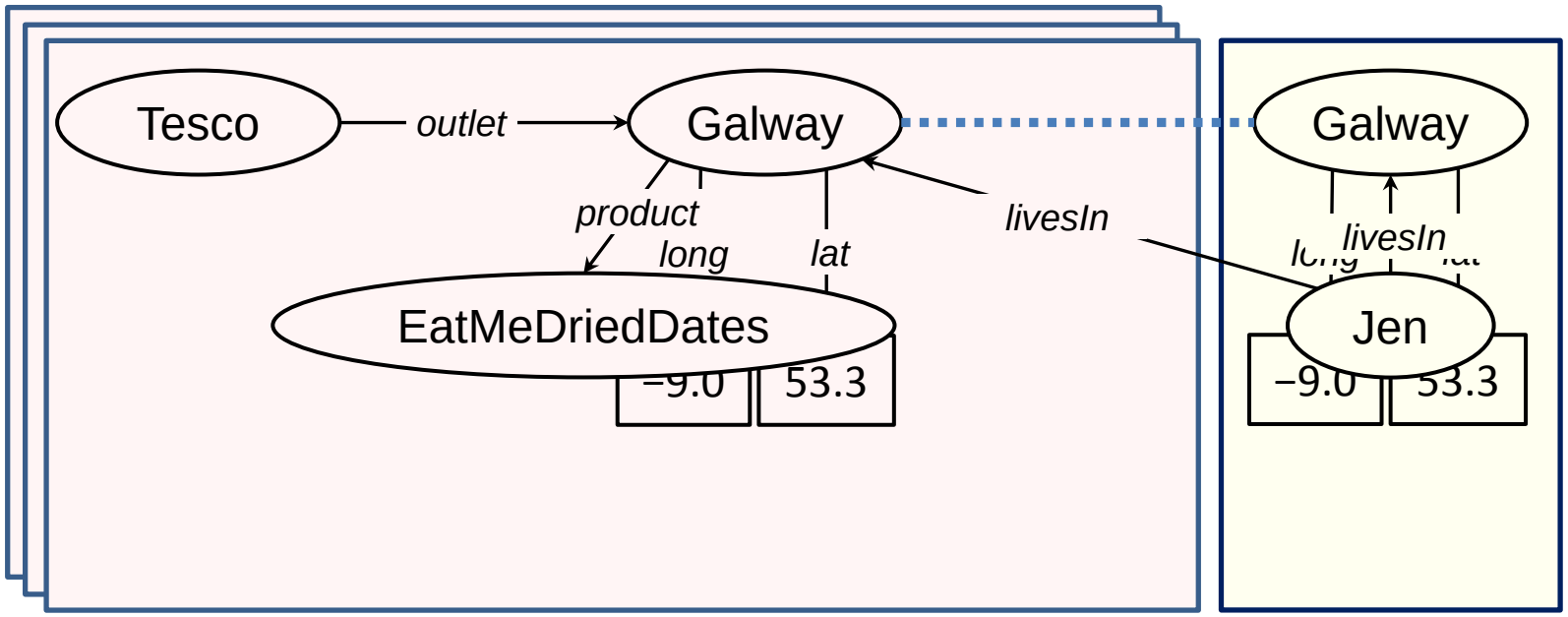
Automatic Integration possible …

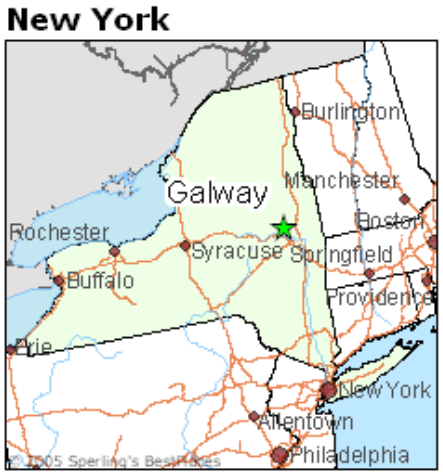


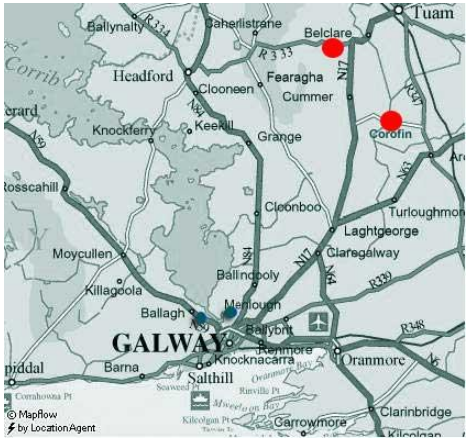
Automatic Integration possible …



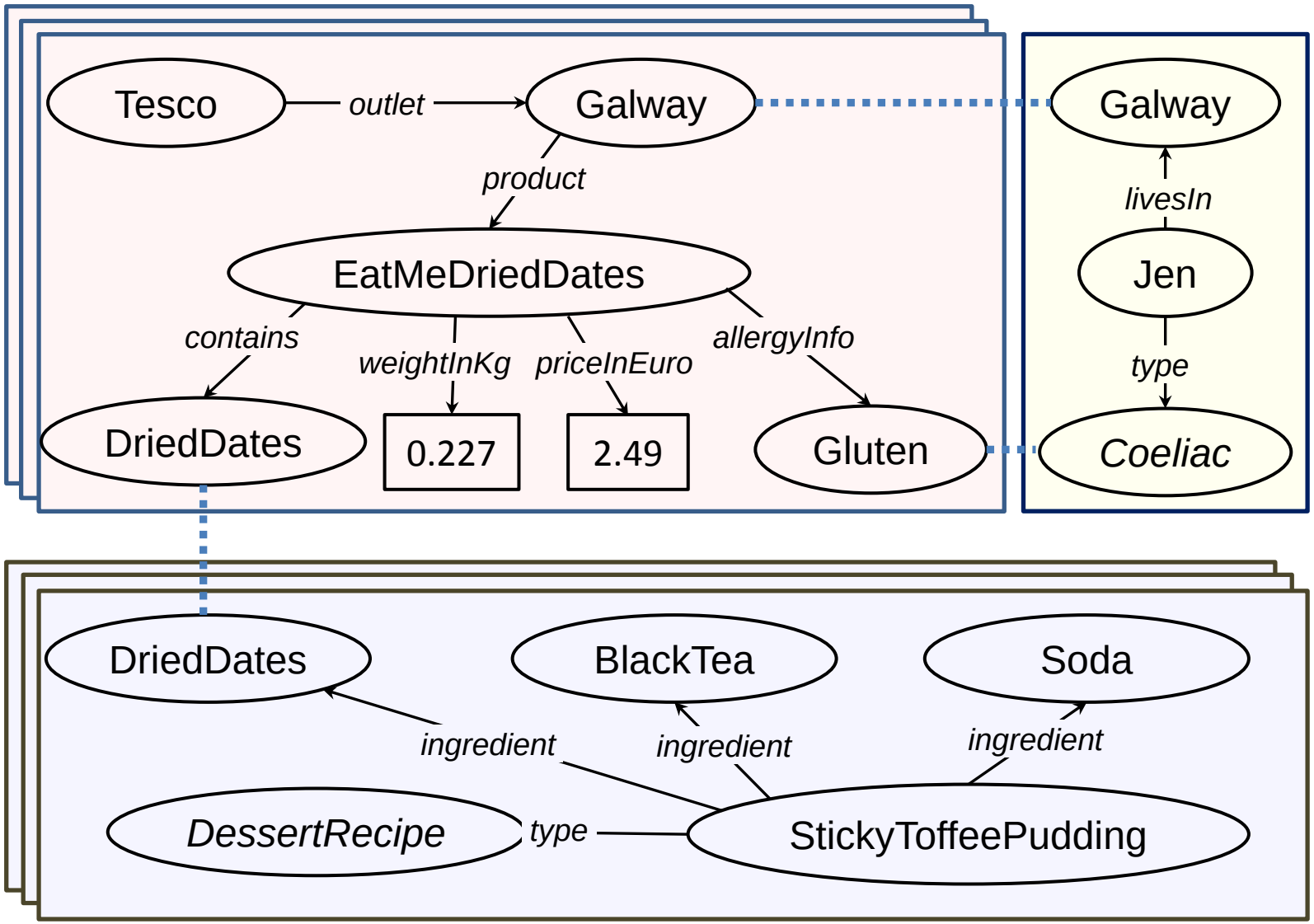
Automatic Integration possible …



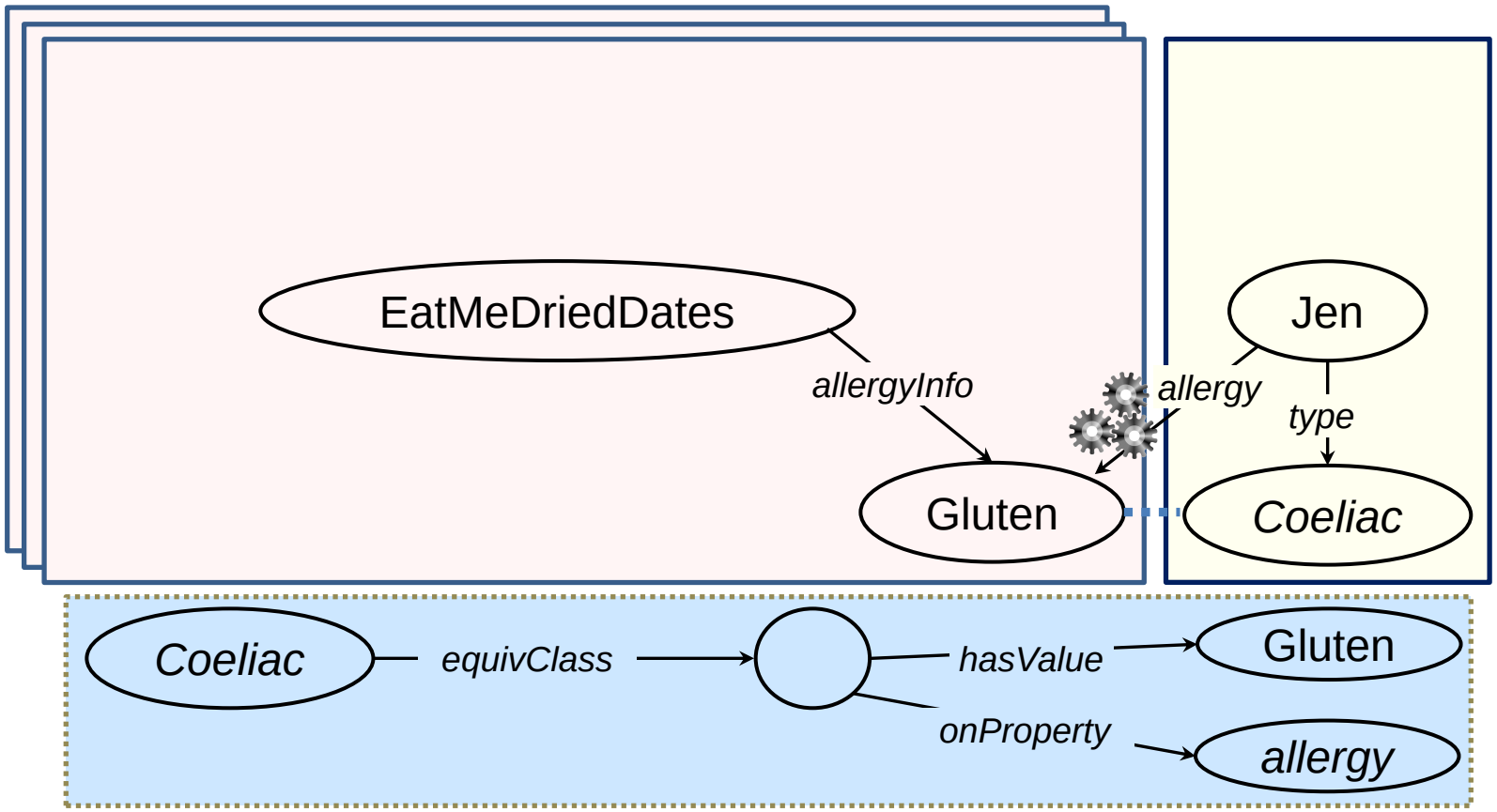




Automatic Integration possible …



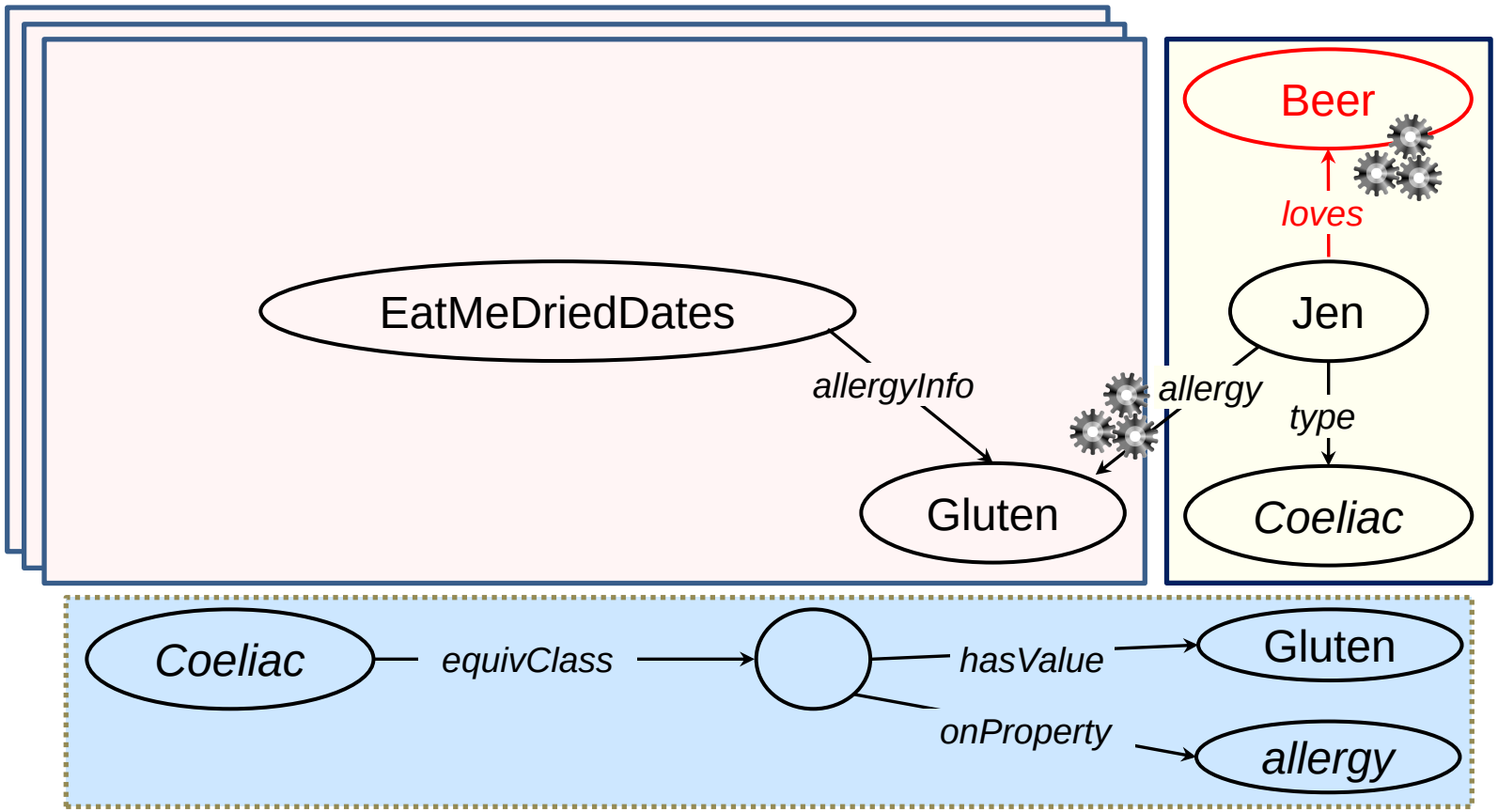
Automatic Integration possible …



Relationship between Coeliac and Gluten?

- Coeliac(Jen) [Jen is a Coeliac]
- Coeliac ≡ ∃allergy.{Gluten} [a Coeliac has allergy to Gluten]
- ∴ allergy(Jen,Gluten) [Jen has allergy to Gluten]

But not everything on the Web is true



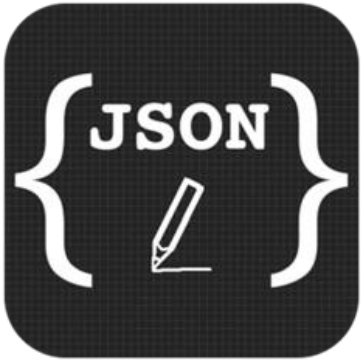
- Coeliac(Jen)
- Coeliac ≡ ∃allergy.{Gluten}
- ∴ allergy(Jen,Gluten)

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

loves beer
loves beer
loves beer

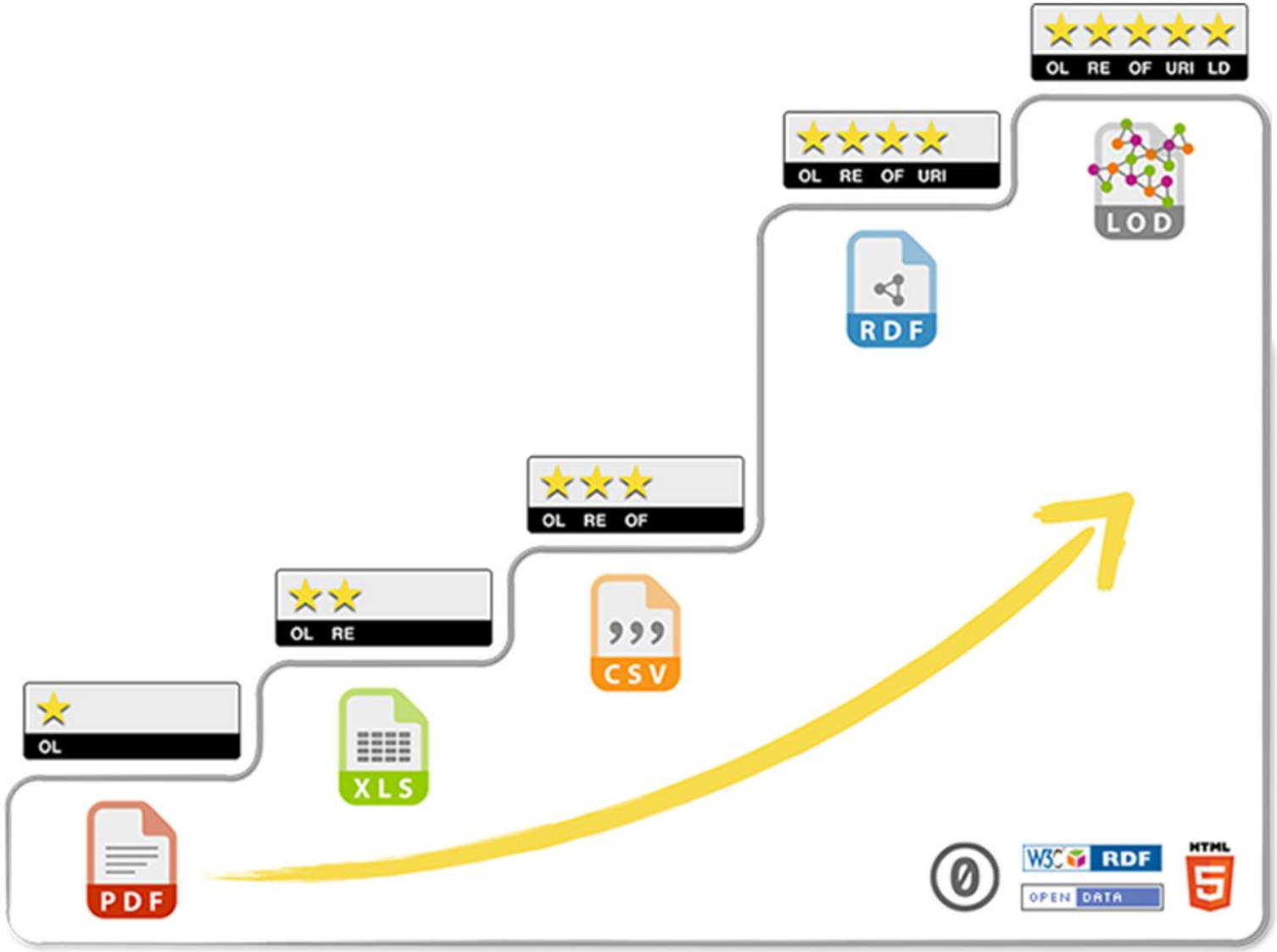
BIG CATCH NUMBER 4: LEGACY DATA NOT RDF

Most Web (meta-)data in …



… and so on …

From ★★★ to ★★★★ is a big step!!



Next week: RDB2RDF

- From relational databases (RDB) to RDF …

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!

RECAP

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:
  - 45% labs
  - 15% project
  - 10% reading group
  - 30% exam

Speculative Schedule ☺ (NOT DEFINITE!)

- Now: Week 11(1)
- 11(2): Lab
- 12(1): RDB2RDF
- 12(2): Lab
- 13(1): Project
- 13(2): Project
- 14(1): Project
- 14(2): Project Presentations
- 15(1): Lecture (Recap/Revision)
- 15(2): Auxiliar / Q&A / No class?

Questions?