SRI International, Artificial Intelligence Center
Menlo Park, USA, 24 July 2009

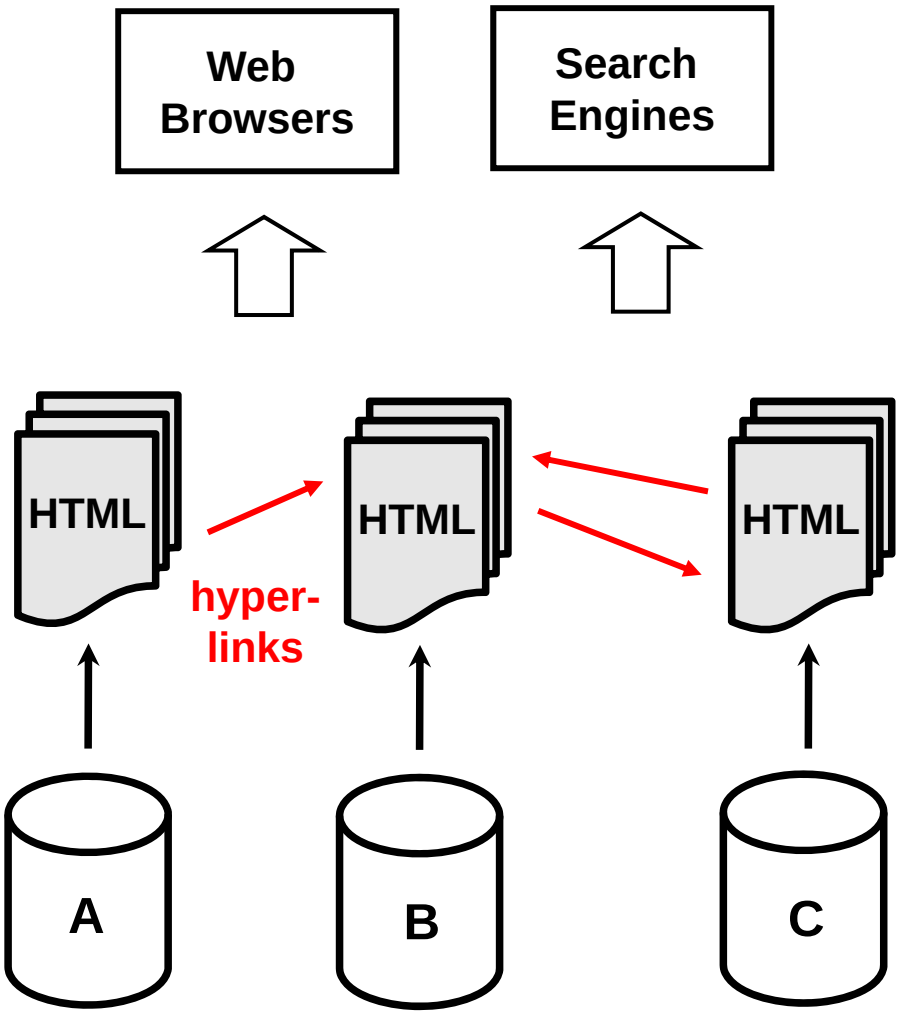
# The Emerging Web of Linked Data

Chris Bizer, Freie Universität Berlin

# Outline

1. **From a Web of Documents to a Web of Data**
   - Web APIs and Linked Data
2. **Linked Data Deployment on the Web**
   - What data is out there?
3. **Applications**
   - What is being done with the data?
4. **Next steps**
   - What is still missing?

# The Classic Web




## Single Global Information Space

1. **URLs as**
   - globally unique IDs
   - retrieval mechanism
2. **HTML as shared content format**
3. **Hyperlinks**

# Problem and Solution

## Problem

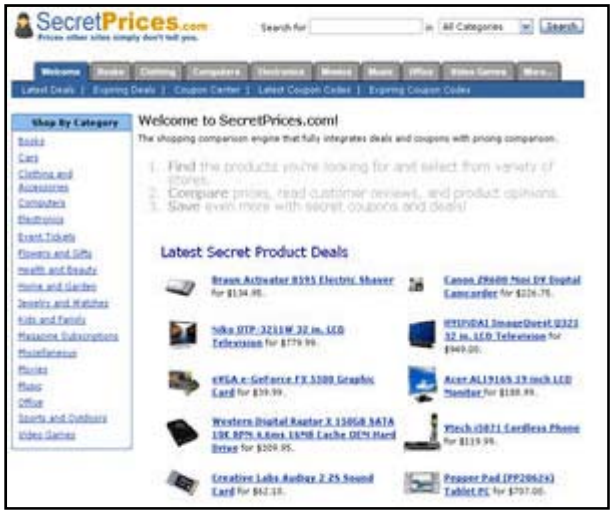
As Web content is only loosely structured it is difficult for applications to do smart things with it.

## Solution

Increase the structure of Web content.

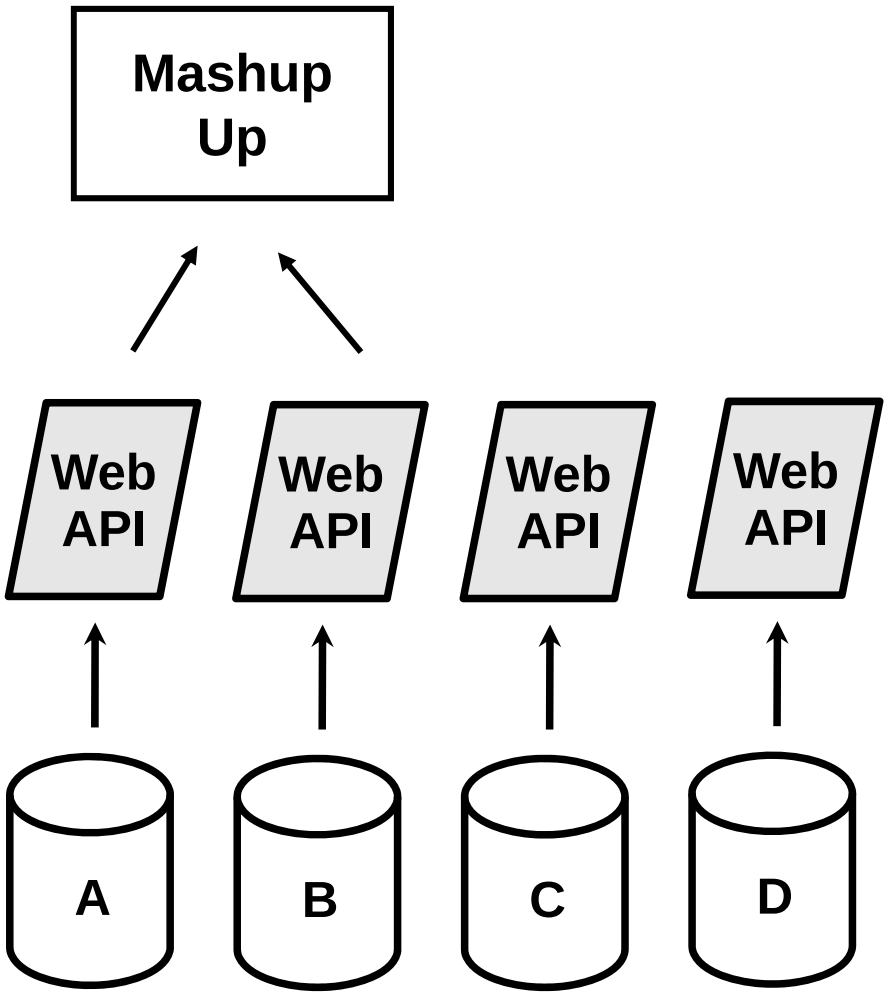
# Web APIs and Mashups

programmableweb



# Web APIs and Mashups

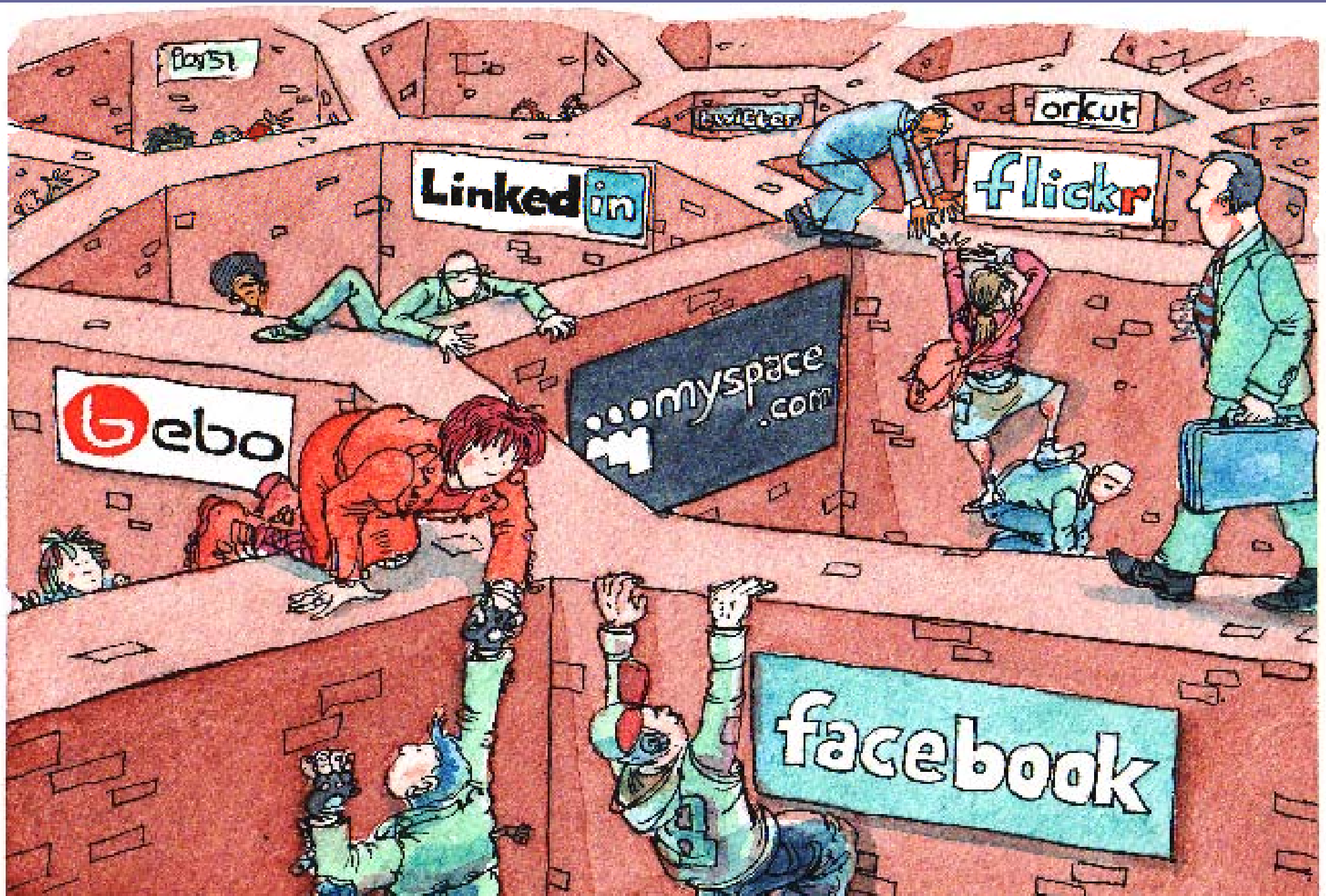



programmableweb

## Shortcomings

1. APIs provide proprietary interfaces
2. Mashups are based on a fixed set of data sources.
3. You can not set hyperlinks between data objects.

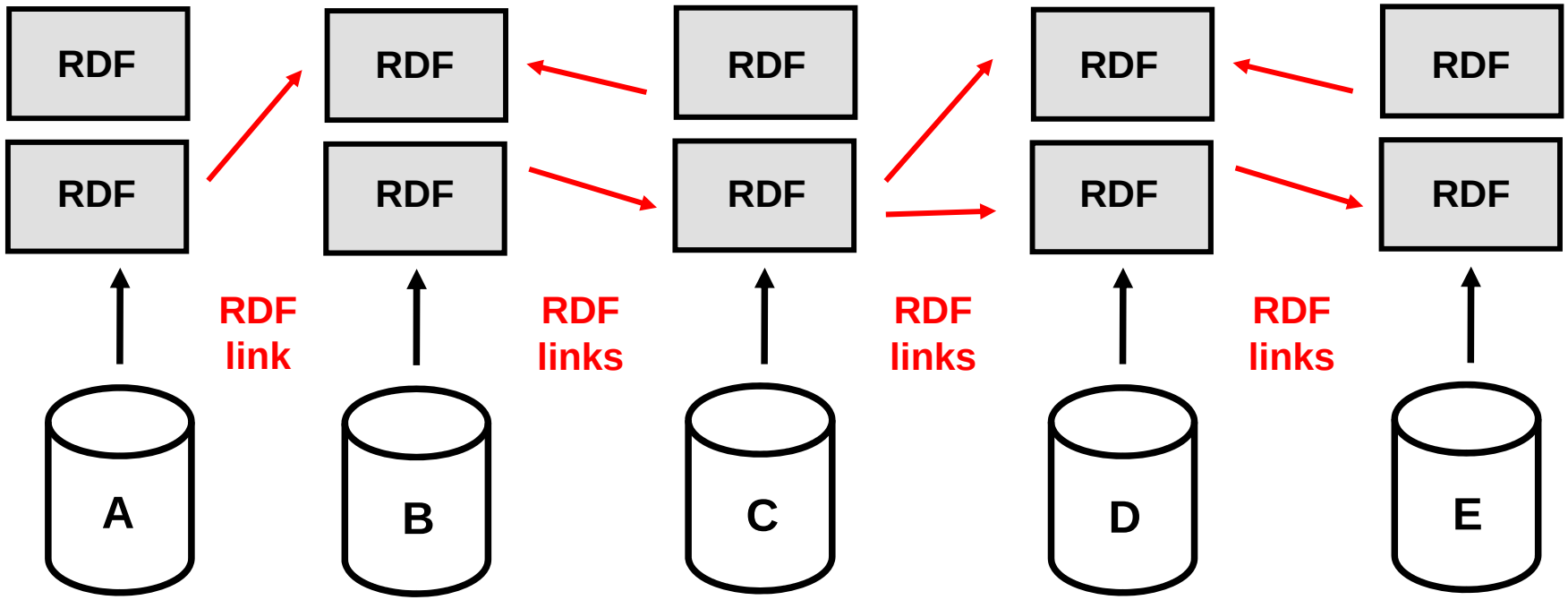
# Web APIs slice the Web into Walled Gardens



# Linked Data

**Use Semantic Web technologies to**

1. **publish structured data on the Web,**
2. **set links between data from one data source to data within other data sources.**



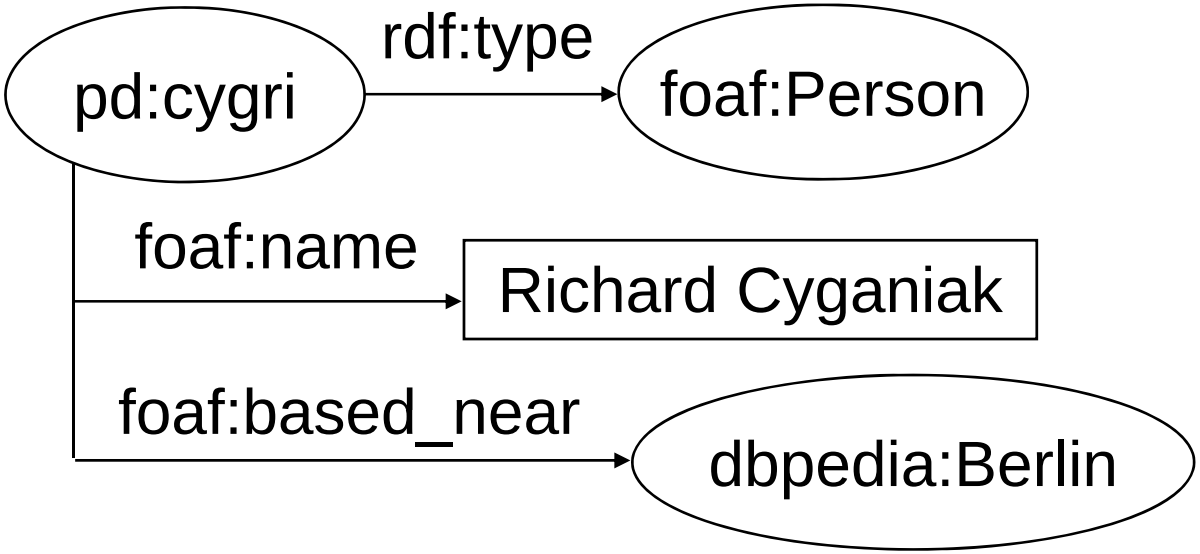
# Linked Data Principles

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 RDF information.**
4. **Include RDF statements that link to other URIs so that they can discover related things.**

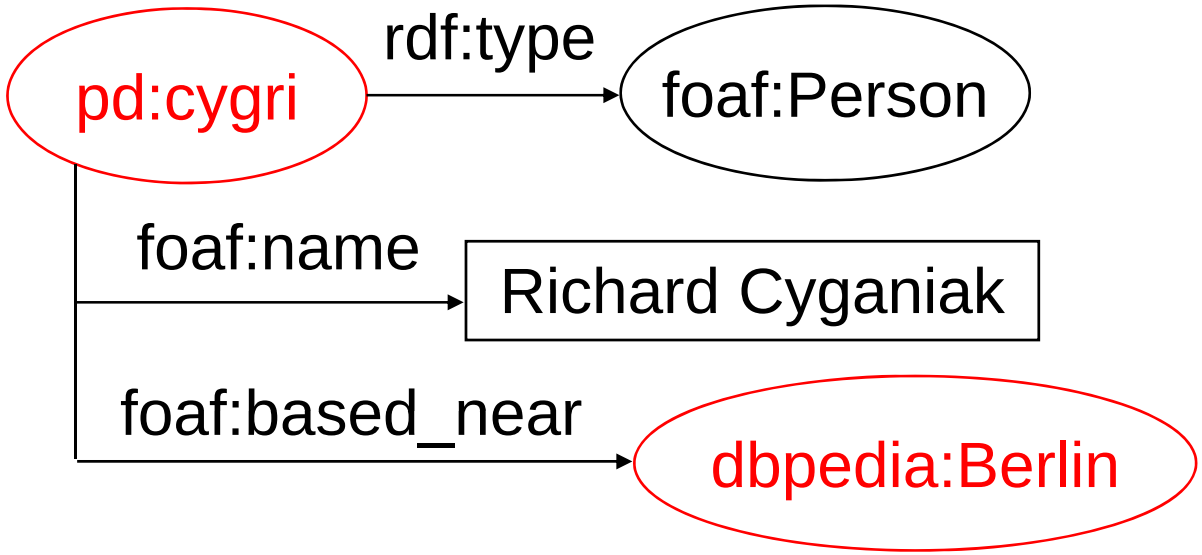
**Tim Berners-Lee 2007**

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

# The RDF Data Model



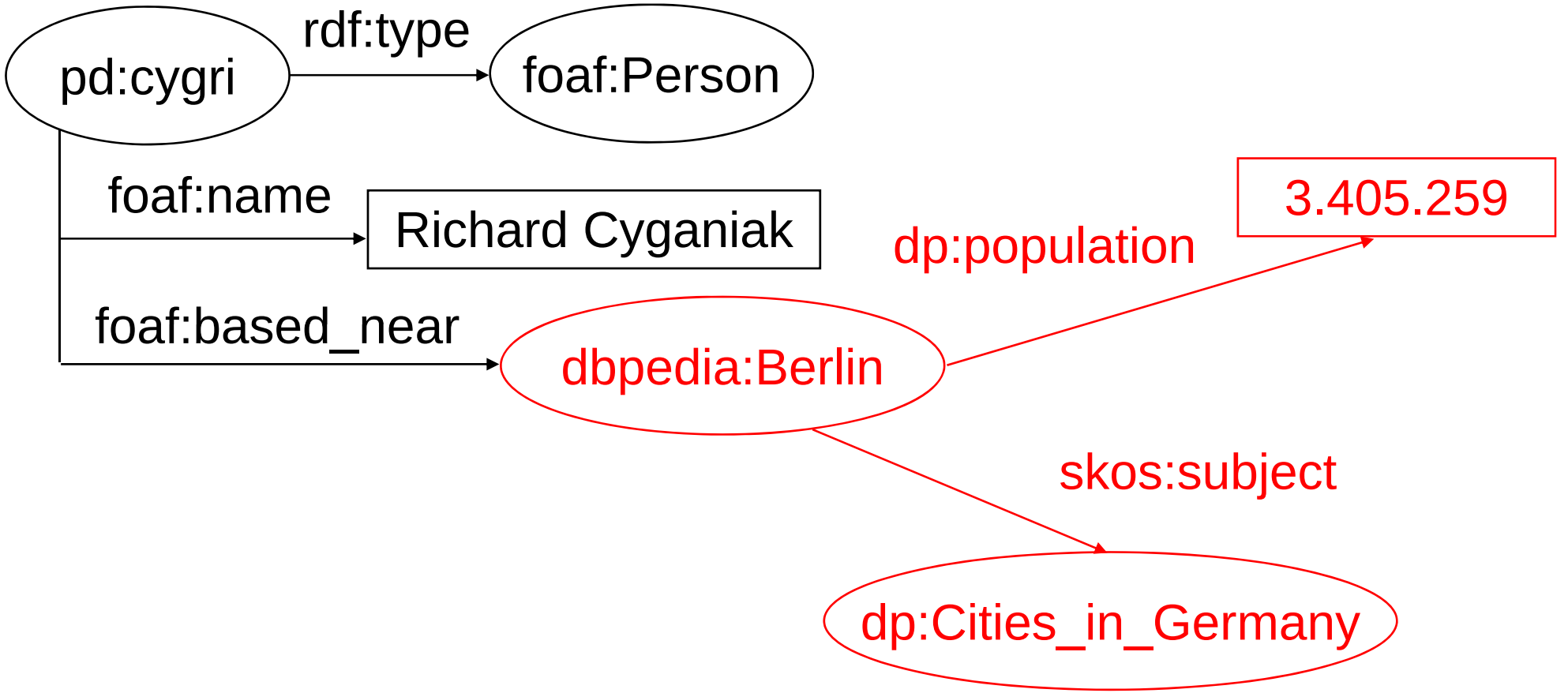
# Data items are identified with HTTP URIs



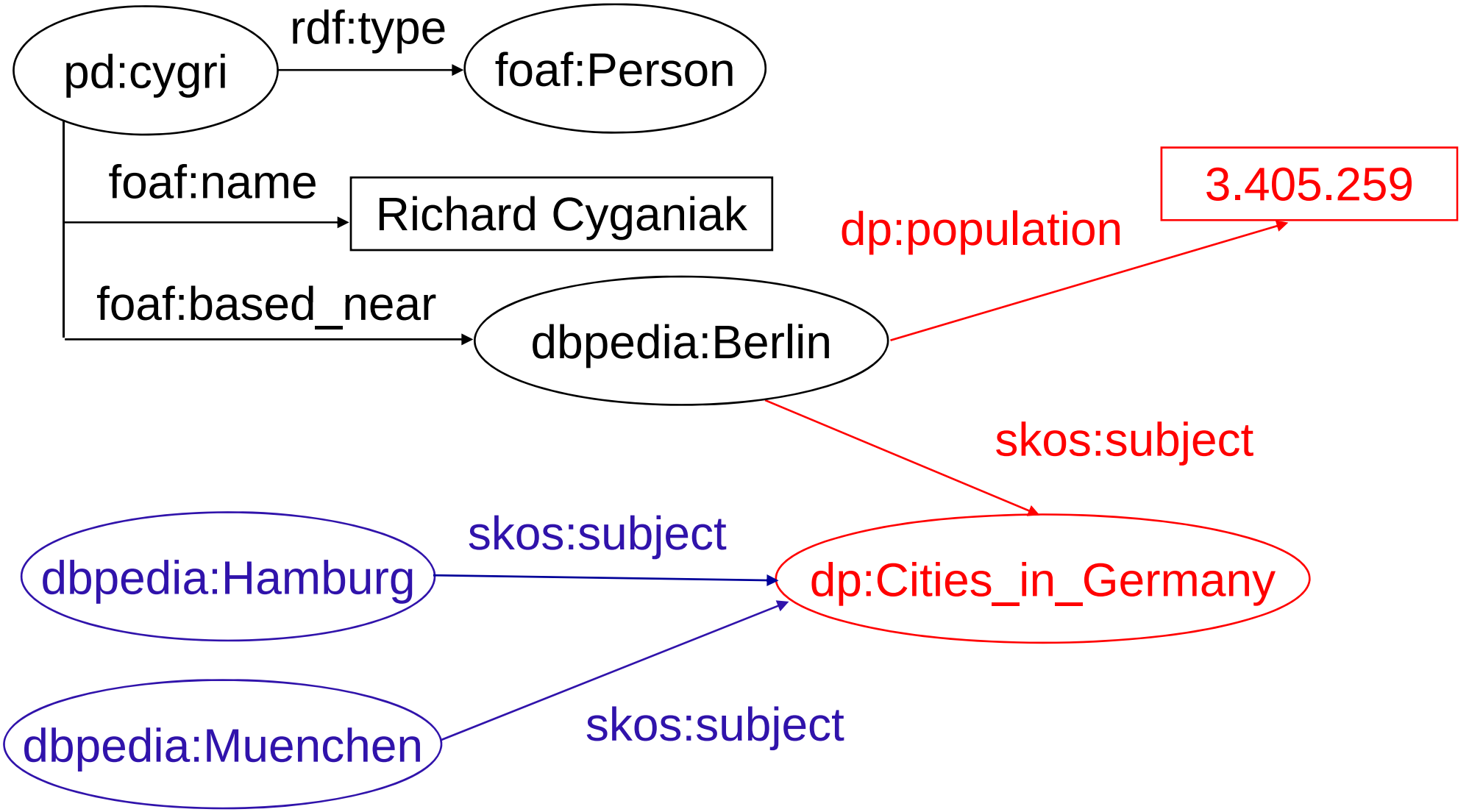

**pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri**

**dbpedia:Berlin = http://dbpedia.org/resource/Berlin**

# Resolving URIs over the Web



# Dereferencing URIs over the Web



Disco - Hyperdata Browser (About)

# Richard Cyganiak

URI: http://richard.cyganiak.de/foaf.rdf#cygri Go!

| Property | Value | Sources |
|---|---|---|
| event | ... | G2 |
| type | http://xmlns.com/foaf/0.1/Person | G1 G2 G3 G4 |
| seeAlso | http://richard.cyganiak.de/cygri.rdf | G2 |
| seeAlso | http://richard.cyganiak.de/foaf.rdf | G3 |
| nearest airport | ... | G1 |
| phone | tel:+49-175-5630408 | G1 |
| sameAs | Richard Cyganiak | G1 |
| based_near | ... | G1 |
| based_near | Berlin | G1 |
| based_near | http://sws.geonames.org/2950159/ | G1 |
| currentProject | http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs | G3 |
| currentProject | http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq | G3 |
| depiction |  | G4 |
| gender | male | G1 |

Disco - Hyperdata Browser (About)

# Berlin

URI: http://dbpedia.org/resource/city/Berlin Go!

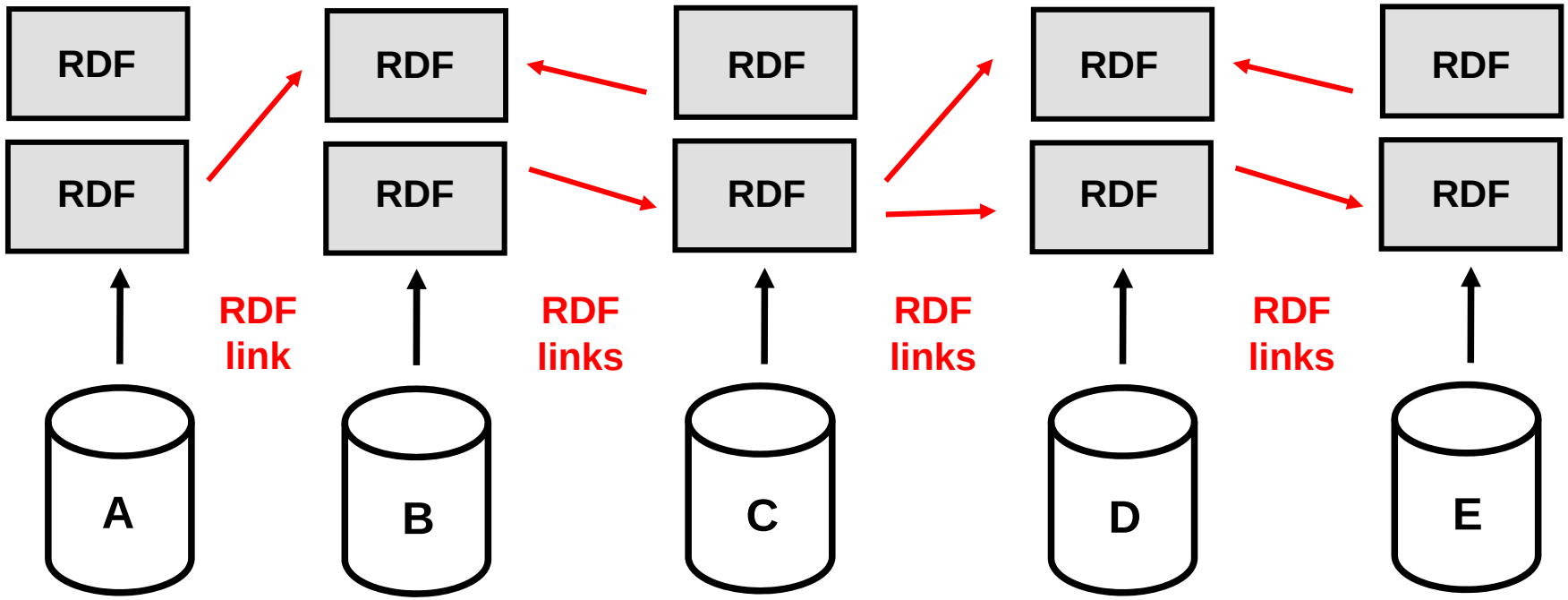
| Property | Value | Sources |
|---|---|---|
| population | 3398888 | G2 |
| type | http://dbpedia.org/City | G2 |
| comment | Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union. | G2 |
| comment | Berlin ist die deutsche Bundeshauptstadt und als Stadtstaat ein eigenständiges Land der Bundesrepublik Deutschland. Berlin ist die bevölkerungsreichste und flächengrößte Stadt Deutschlands und nach Einwohnern die zweitgrößte Stadt der EU. | G2 |
| label | Berlin | G2 |
| sameAs | http://sws.geonames.org/2950159/ | G2 |
| subject | http://dbpedia.org/resource/category/Berlin | G2 |
| subject | http://dbpedia.org/resource/category/Capitals_in_Europe | G2 |
| subject | http://dbpedia.org/resource/category/Cities_in_Germany | G2 |
| subject | http://dbpedia.org/resource/category/German_state_capitals | G2 |
| subject | http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games | G2 |
| subject | http://dbpedia.org/resource/category/States_of_Germany | G2 |
| sourceURL | Berlin | G1 |
| depiction |  | G2 |
| page | http://en.wikipedia.org/wiki/Berlin | G2 |
| is birthplace of | Adolf von Baeyer | G2 |

# Properties of the Web of Linked Data

- **Anyone can publish data to the Web of Linked Data**
- **Entities are connected by links**
  - creating a global data graph that spans data sources and enables the discovery of new data sources.
- **Data is self-describing**
  - If an application encounters data represented using an unfamiliar vocabulary, the application can resolve the URIs that identify vocabulary terms in order to find their RDFS or OWL definition.
- **The Web of Data is open**
  - meaning that applications can discover new data sources at run-time by following links.

# 2. Linked Data Deployment on the Web

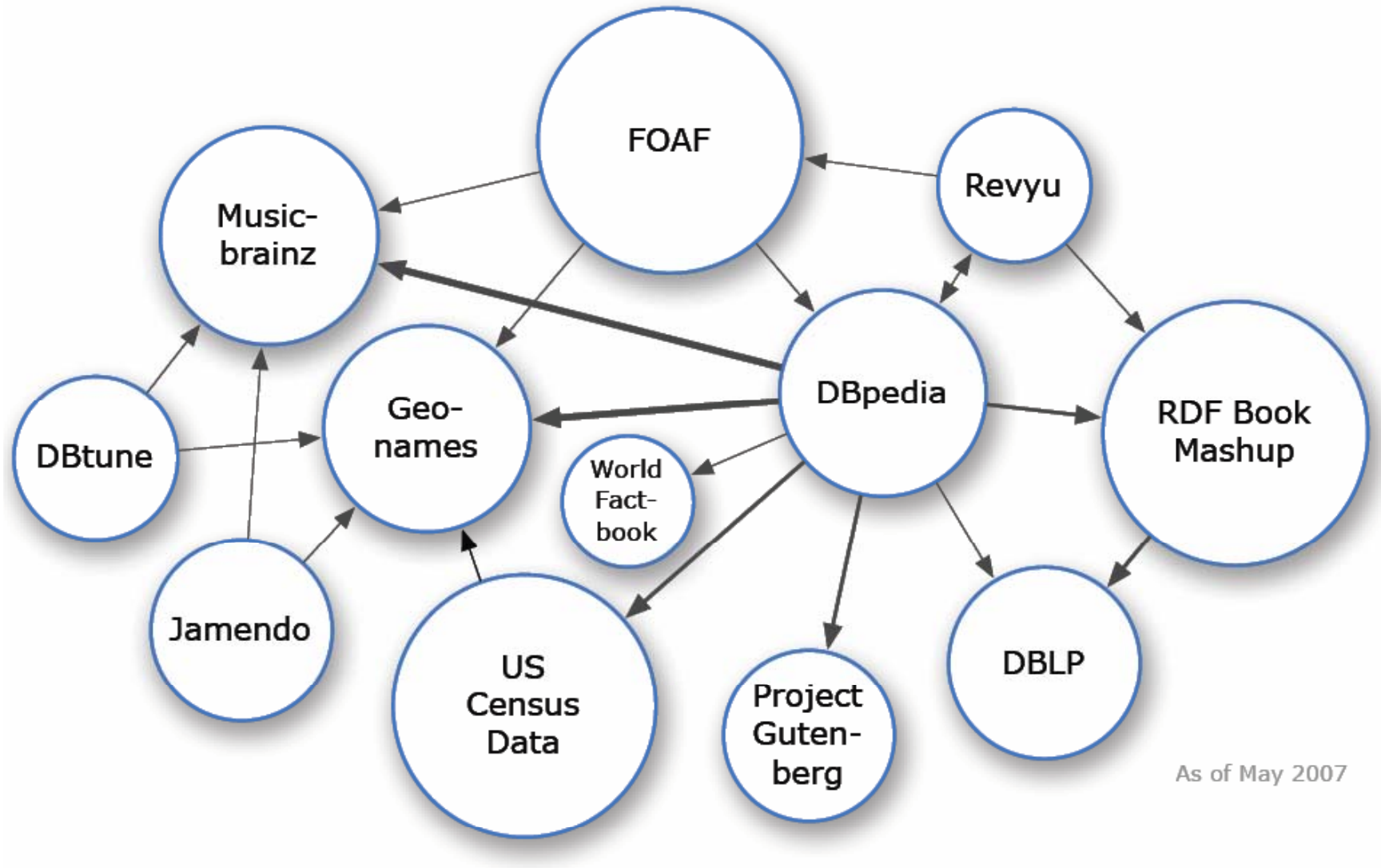
- Is this real?



# W3C Linking Open Data Project


- **Grassroots community effort to**
  - publish existing open license datasets as Linked Data on the Web
  - interlink things between different data sources

# LOD Datasets on the Web: May 2007




- Over 500 million RDF triples
- Around 120,000 RDF links between data sources

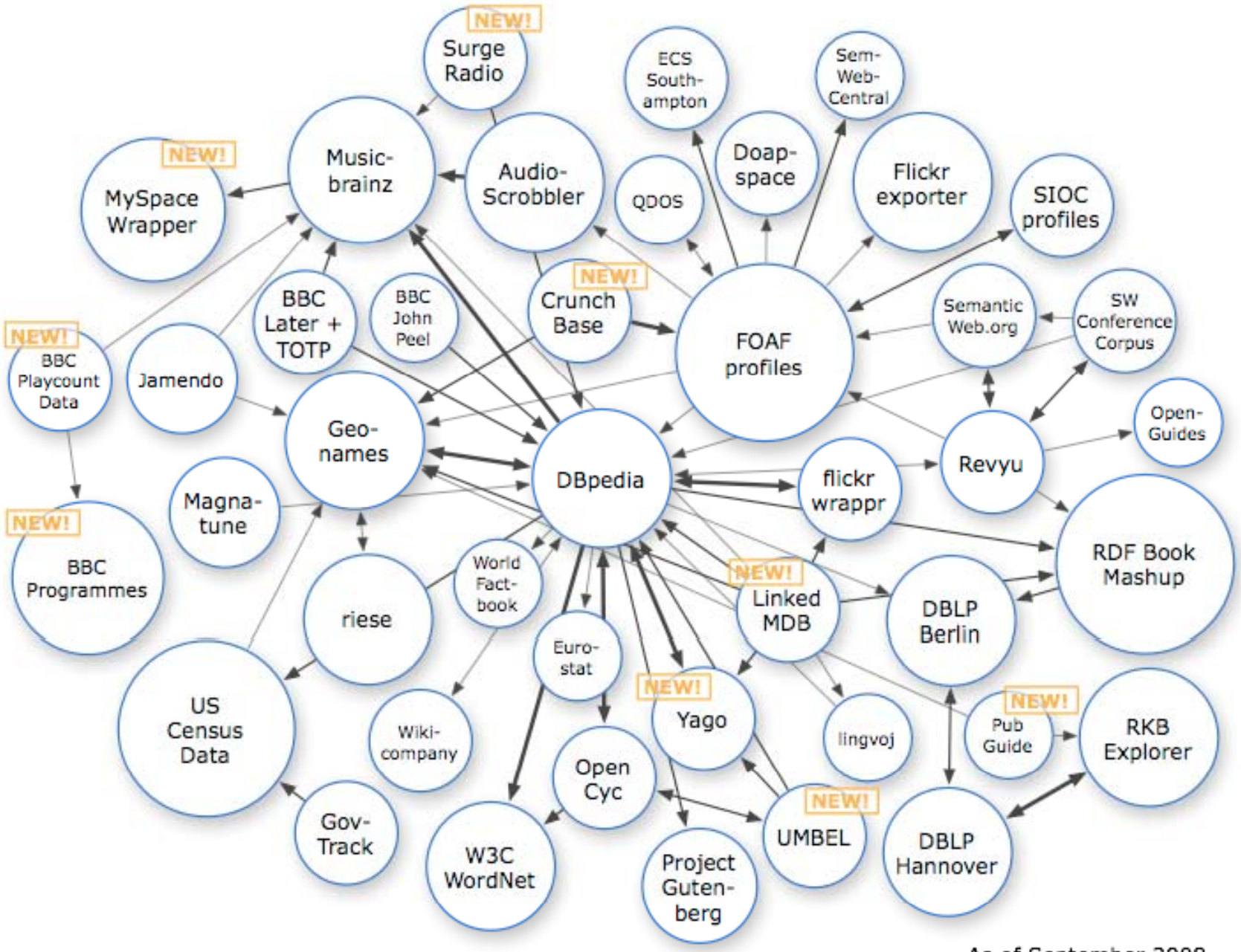
# Example RDF Links

- RDF links from DBpedia to other data sources

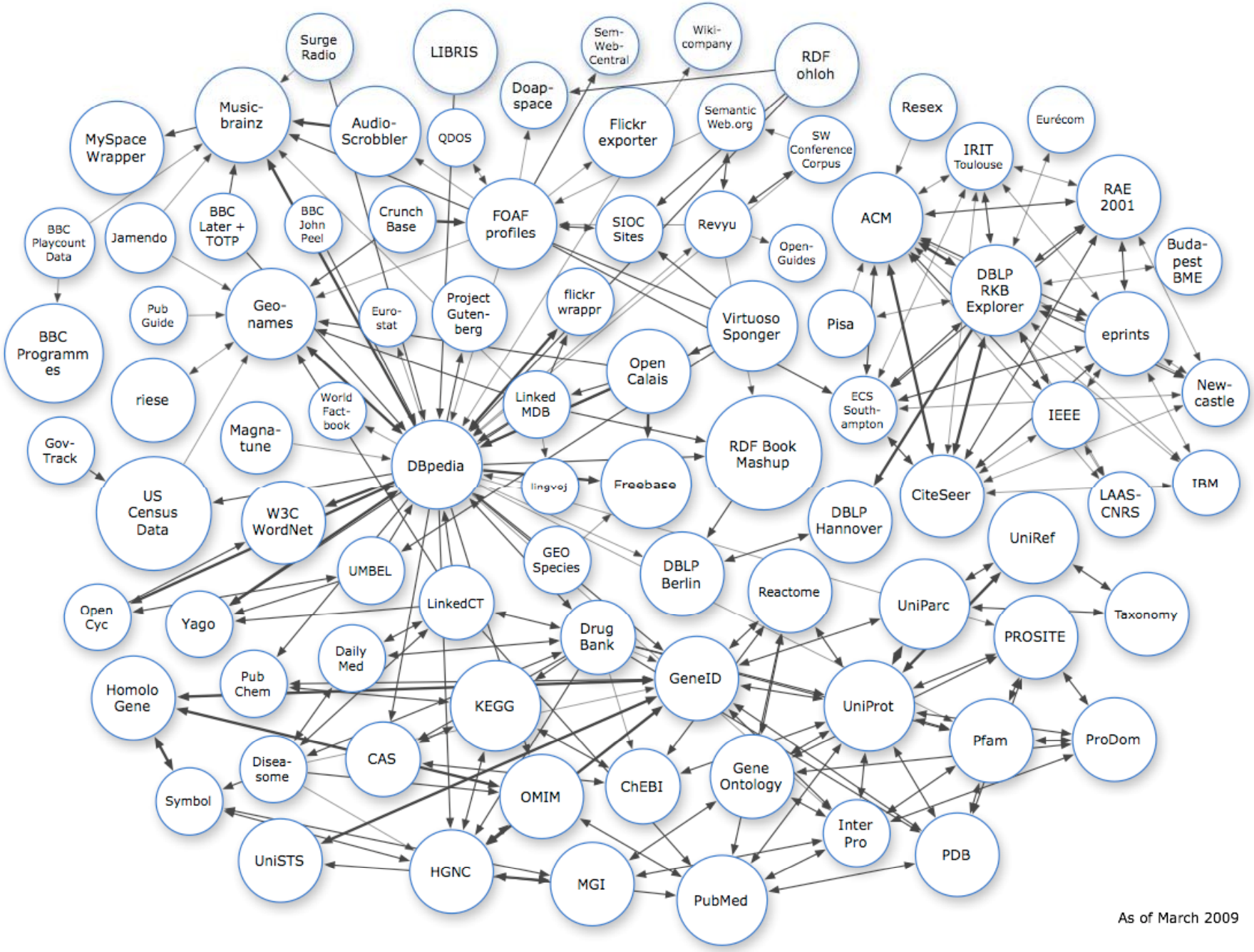
```
<http://dbpedia.org/resource/Berlin> owl:sameAs
<http://sws.geonames.org/2950159> .
```

```
<http://dbpedia.org/resource/Tim_Berners-Lee> owl:sameAs
<http://www4.wiwiss.fu-berlin.de/dblp/resource/person/100007> .
```

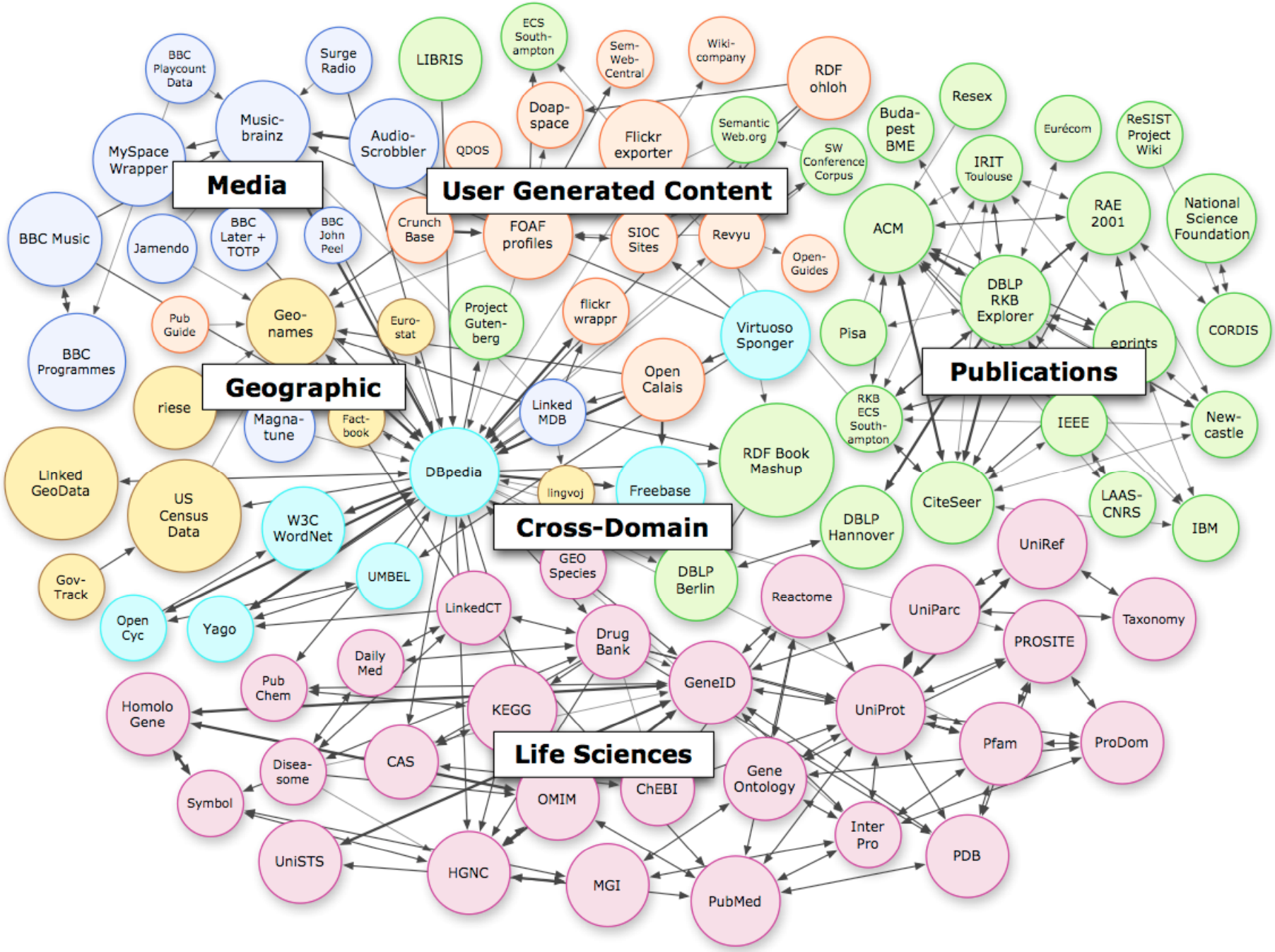
# LOD Datasets on the Web: September 2008



# LOD Datasets on the Web: March 2009



# LOD Datasets on the Web: July 2009



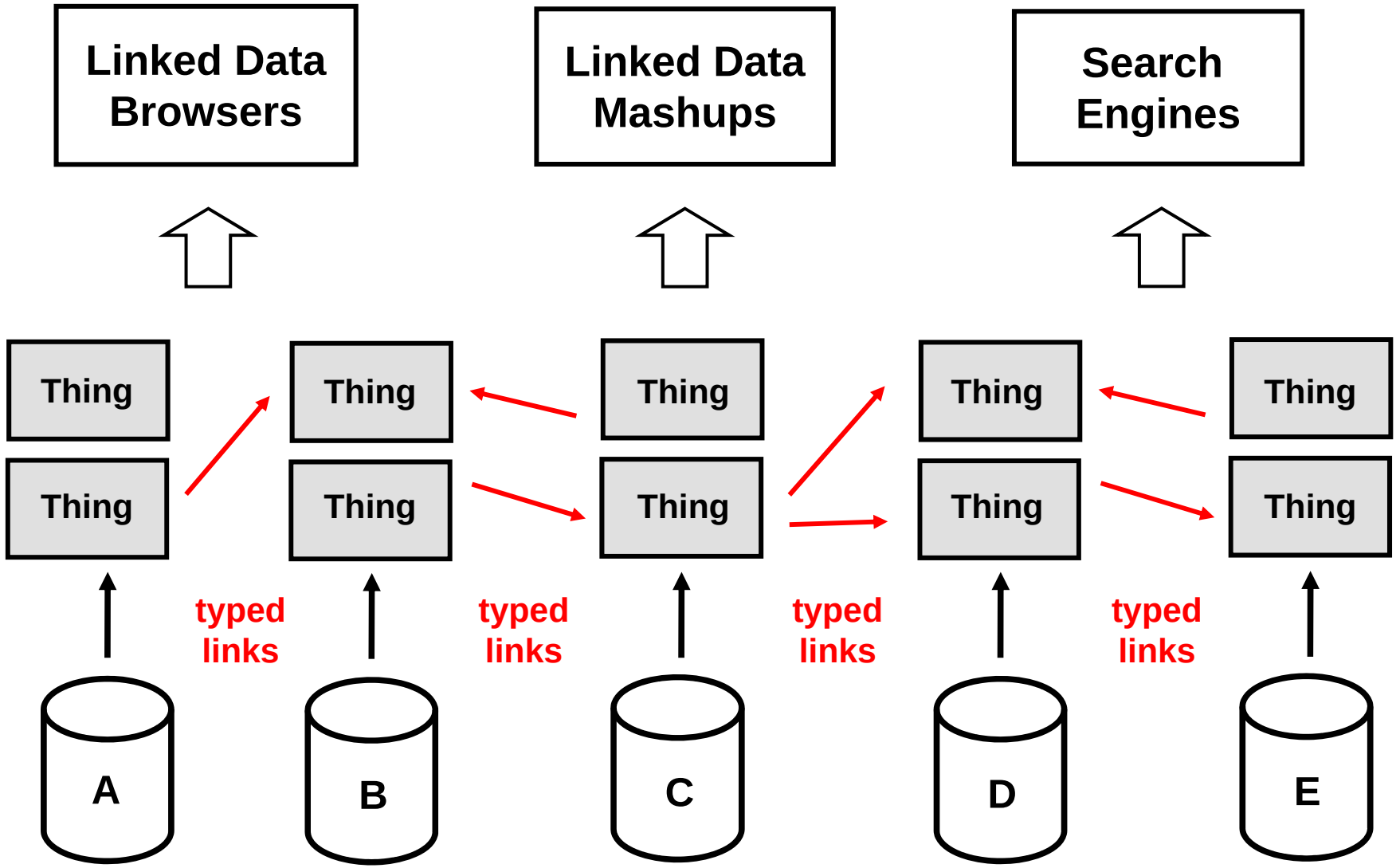
# LOD data set statistics as of July 2009

| *Domain* | *No of Triples* | *% of Cloud* | *No of Links* | *% of Links* |
|---|---|---|---|---|
| Media | 698.000.000 | 10,4% | 1.238.000 | 0,8% |
| Publications | 212.000.000 | 3,2% | 4.922.000 | 3,3% |
| Life Sciences | 2.429.000.000 | 36,1% | 133.199.000 | 89,4% |
| Geographic Data | 3.097.000.000 | 46,0% | 4.038.000 | 2,7% |
| User Generate Content | 76.000.000 | 1,1% | 1.559.000 | 1,0% |
| Cross-Domain | 214.000.000 | 3,2% | 3.992.000 | 2,7% |
| ***Total*** | ***6.726.000.000*** | | ***148.948.000*** | |

**+ 2 billion triples from Data.gov published yesterday.**

# 3. Applications

- What can I do with this?



# Linked Data Browsers

- Tabulator Browser (MIT, USA)
- Marbles (FU Berlin, DE)
- OpenLink RDF Browser (OpenLink, UK)
- Zitgist RDF Browser (Zitgist, USA)
- Humboldt (HP Labs, UK)
- Disco Hyperdata Browser (FU Berlin, DE)
- Fenfire (DERI, Irland)

http://www.w3.org/People/Berners-Lee/card#i Open

marbles

# Tim Berners-Lee

| | |
|---|---|
| http://www.w3.org/1999/02/22-rdf-syntax-ns#type | • Person<br>• http://www.w3.org/2000/10/swap/pim/contact#Male |
| label | • Tim Berners-Lee |
| sameAs | • Tim Berners-Lee (also at www4.wiwiss.fu-berlin.de) |
| image |  |
| Weblinks | http://www.w3.org/People/Berners-Lee/ |
| name | • Tim Berners-Lee<br>• Timothy Berners-Lee<br>• Tim Berners Lee |
| Given name | • Timothy |
| family_name | • Berners-Lee |
| sha1sum of a personal mailbox URI name | • 965c47c5a70db7407210cef6e4e6f5374a525c5c |
| workplace homepage | • http://www.w3.org/ |
| nickname | • TimBL |
| nickname | • TimBL<br>• timbl |
| personal mailbox | • mailto:timbl@w3.org |
| seeAlso | • Tim Berners-Lee's FOAF file<br>• Tim Berners-Lee's FOAF file |
| is seeAlso of | • Tim Berners-Lee |

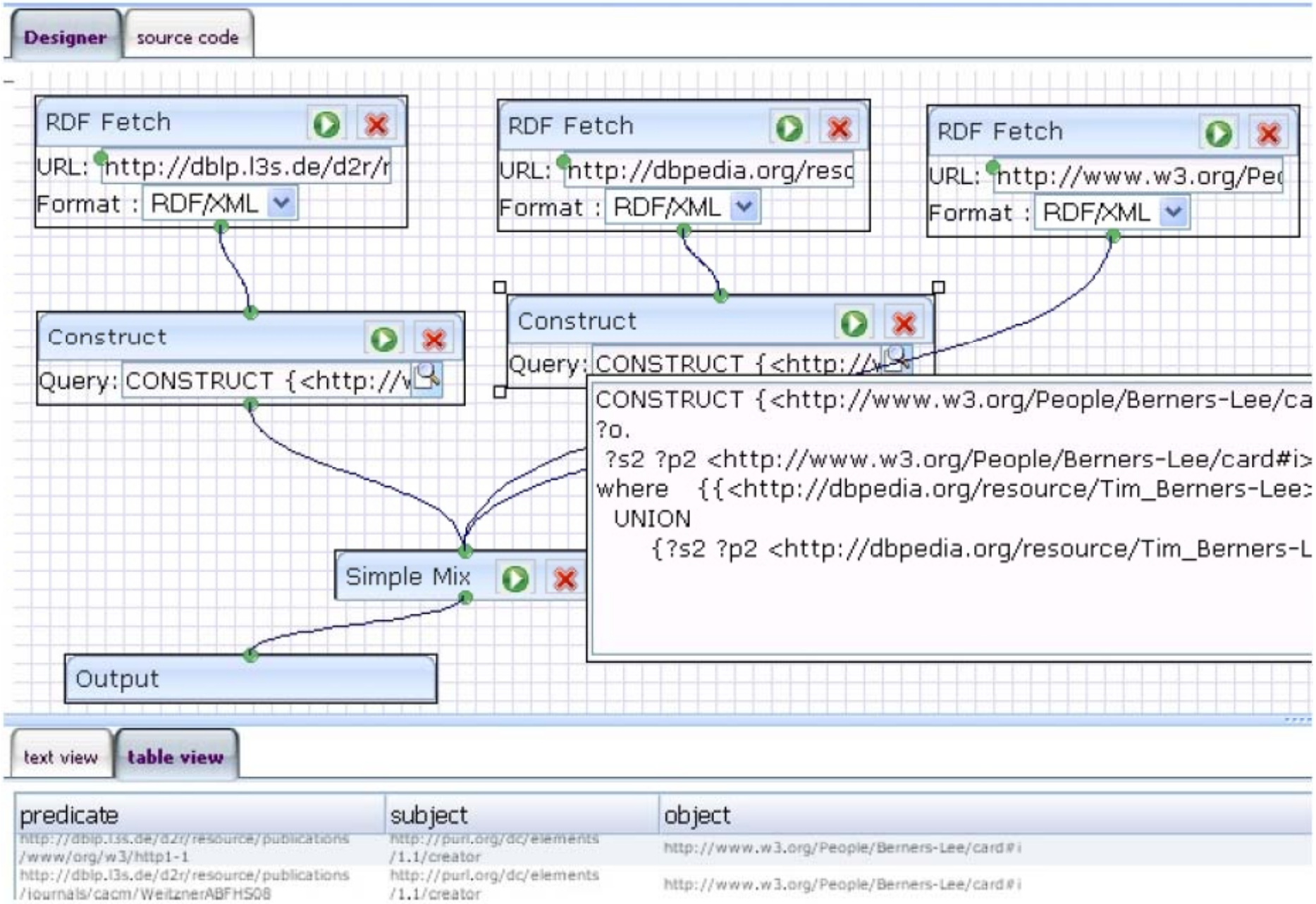
# Linked Data Mashups

- Domain-specific applications using Linked Data from the Web

# DBpedia Mobile


- Geospatial entry point into the Web of Data
- Starts with DBpedia, Revyu and Flickr data

# DERI Semantic Web Pipes



# Web of Data Search Engines

- Falcons (IWS, China)
- Sig.ma (DERI, Ireland)
- Swoogle (UMBC, USA)
- VisiNav (DERI, Ireland)
- Watson (Open University, UK)

Falcons

Object Search Concept Search

Beijing Search Objects

Supports Boolean operators, quotes, and wildcard characters.

All

**Artifact** Capital City City Document Group
Institution Landmark Location Noun Synset Ontology
Organization Person Publication Subject System

Objects 1 - 10 of 8634 for your search **Beijing** (1.223 seconds)

**Beijing**
**Types**: Capital, City
**Labels**: 北京" || Pekin || Пекин" || 北京市" || Pequim || Pechino || **Beijing** || Pékin" || Peking || Pekín"
http://dbpedia.org/resource/**Beijing** - Described in 184 documents

**Beijing**
**Types**: Subject,
**Labels**: **Beijing**
http://ontoworld.org/wiki/Special:URIResolver/**Beijing** - Described in 11 documents

**Beijing** Guoan
**Types**: Club
**Labels**: **Beijing** Hyundai || 北京国安" || 北京国安足球俱乐部" || **Beijing** Guoan
http://dbpedia.org/resource/**Beijing**_Guoan - Described in 30 documents

**Beijing**
The declaration of this URI may be unauthorized.
**Types**: Capital City
**Labels**: **Beijing**
http://lonely.org/russia#**Beijing** - Described in 5 documents

VisiNav

Tim Berners-Lee ☒ Knows ☒ weblog ☒ New Search Ok

Detail View List View Table View Timeline View RSS next ▶ Results 1 - 10 of 54

## Ivan Herman
http://www.ivan-herman.net/
Document Resource Document

## breadcrumbs
http://dig.csail.mit.edu/breadcrumbs/blog/2
RSS1.0 News Channel Document Resource

## Ivan's private site
http://ivan-herman.name/
RSS1.0 News Channel Document Resource

## open source
http://www.advogato.org/person/connolly/
RSS1.0 News Channel Document organization
Advogato blog for connolly
2009-05-31T20:23:14Z

## Paul Downey
http://blog.whatfettle.com/
Document Resource Document

SIG.MA
SEMANTIC INFORI
MASHUP

Help | About | Forum

Chris Bizer

Add More Info | Start New | Options | Order | Permalink

# Chris Bizer

**picture:**  [3]  [5] 

[16]

**given name:** Chris [3,5,9,10,16]

**family name:** Bizer [3,5,9,10,16]

**is creator of:**
- DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food [6,18]
- http://data.semanticweb.org/conference/eswc/2007/demo-3 [9]
- The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]
- D2R Server - Publishing Releational Databases on the Semantic Web. [16]
- Named Graphs, Provenance and Trust [16]
  - hide value | just this value | which sources | reject sources [6]
  - RAP: RDF API for PHP [16]
- Fresnel: A Browser-Independent Presentation Vocabulary for RDF [16]

Sources (20) | Appro

1 Chris Bizer - Free Univ
http://videolectures.net/chris_

2 Chris Bizer - semantic
http://ontoworld.org/wiki/Chris

3 Untitled document 6 f
BOSS http://www.faceboo

4 Chris Bizer - semantic
http://semanticweb.org/wiki/Ch

5 Chris Bizer - LinkedIn
BOSS http://www.linkedin.

6 Chris Bizer 10 facts | 20
http://data.semanticweb.org/p

7 Chris Bizer - semantic
http://semanticweb.org/index.p

8 Flickr: Chris Bizer's Ph
BOSS http://flickr.com/pho

9 Untitled document 8 f
http://data.semanticweb.org/c

10 Chris Bizer 6 facts | 20
BOSS http://ebiquity.umbc

<- 1 2 ->

http://example.loc/doc

# What are the big players doing?

- **Yahoo! and Google have started to crawl Linked Data in its RDFa serialization as well as Microformats.**
- **Yahoo!**
  - provides access to crawled data through the Yahoo BOSS API
  - is using the data within Yahoo Search Monkey to make search results more useful and visually appealing.
- **Google**
  - uses crawled RDF data for its Social Graph API
  - is planning to / uses crawled data to enhance search results snippets for reviews and people.

# Yahoo! Search Monkey

# Connecting the classic Web and Linked Data

- **Annotate Web documents with Linked Data URIs**

```
<http://data.semanticweb.org/conference/eswc/2007/paper-69>
dc:subject <http://dbpedia.org/resource/Machine_learning> .
```

- **(Semi-) Automated Annotation Services using Named Entity Recognition**
  - Open Calais (Thomsons Reuters) for news
  - Zemanta (startup) for blog posts

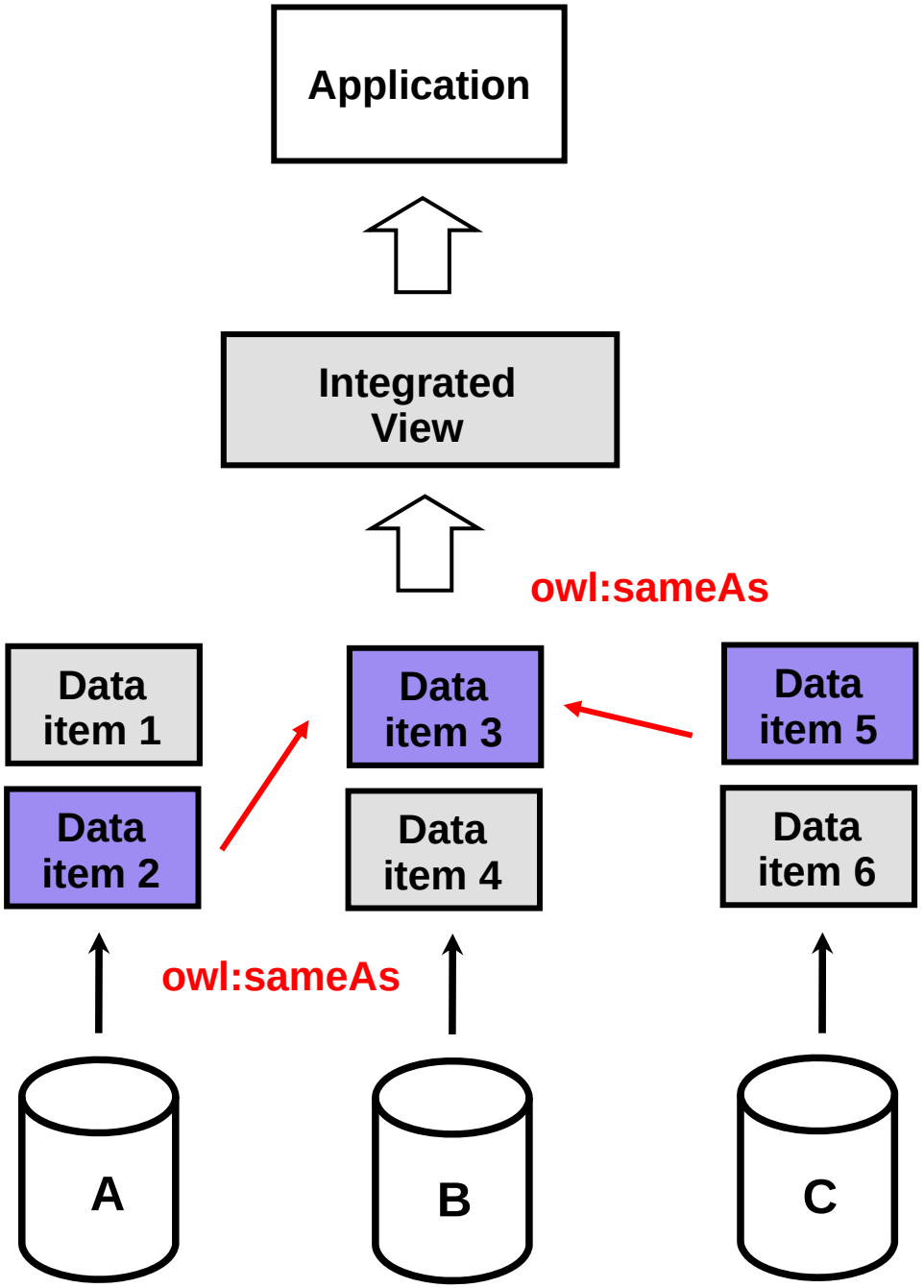


- **Goals**
  - Connect everything.
  - Improve search by using Linked Data as background knowledge.
  - Display Web of Data content as info boxes next to news, blog posts.

# Next steps

- **More data is becoming available ....**
  - US and UK government data
  - bibliographic data via Open Archives ORE
- **What is still missing?**

# Linked Data Fusion




**Applications want an integrated view on all data that is available about an real-world entity!**

# Linked Data Fusion - Requirements

1. **Map data into a single schema**
   - so that data can be rendered and queried properly.
2. **Smush data from all sources about a single real-world entity**
   - while keeping track of information provenance.
3. **Resolve inconsistencies in the data**
   - by applying different trust heuristics.

# Data Quality and Trust

- **There are no facts on the Web!**
- **The Web is a social thing and everything on the Web is a claim.**
- **Therefore we more research on quality assessment, trust, data-cleansing.**
- **Move the trust layer down in the Semantic Web Layer Cake**
  - Right above RDF and below OWL, SPARQL and RIF?

# Reasoning with Linked Data

- Topic at recent Dagstuhl perspectives seminar
- Vocabulary term cherry-picking
- Retrieving ontology fragments from the Web
- Retrieving (partial) mappings from the Web
- Reasoning with large amounts of (inconsistent) Web data

# Pay As You Go Data Integration

- **There is a pay as you go data integration paradigm emerging on the Web of Data**
  - Publish data first using different schemata
  - Maybe use common vocabularies
  - Publish mappings to the Web afterwards

- **How to derive best-effort answers based on heterogeneous Web data and partial mappings?**
  - Alon Halevy, et al.: Web-scale Data Integration: You can only afford to Pay As You Go

# User Interfaces and Interaction Paradigms

- **How do we build interfaces that operate over such large amounts of data?**
  - How to aggregate the data in a meaningfull way?
- **What will be their interaction paradigm?**
  - Will the browser be something like a Web-Excel including drill-down?
  - Will end-users notice that they are using Linked Data?
- **How to explain data provenance and data fusion?**
  - Tim Berner-Lee‘s „Oh, yeah?“ button.
- **What will Google and Yahoo do with the data?**
  - Will search engines turn into answer engines?

# Hands on: How to publish Linked Data?

- **Read the "How to Publish Linked Data on the Web" tutorial**
  - http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
- **Publishing Tools**
  - D2R Server: Publishes relational data bases as Linked Data and via SPARQL
  - Pubby: Linked Data wrapper that can be used together with any RDF store
- **Link Generation Tools**
  - Silk – Link Discovery Framework
  - ODDlinker
- **Join the W3C Linking Open Data community**
  - Wiki: http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
  - Mailing list: public-lod@w3.org

# Thanks!

## References

- Overview Article
  Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data – The Story So Far
  http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf
- Linking Open Data Project Wiki
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
- Tutorial on How to Publish Linked Data on the Web
  http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/