

Introduction to the Semantic Web
Invited Tutorial at ISWC 2010
November 7th, 2010, Shanghai, China

Linked Data

Christian Bizer
Freie Universität Berlin
Germany

1. Foundations of Linked Data

- What is the vision and goal?

2. The Web of Linked Data

- What data is out there?
- What is being done with the data?

3. How to publish Linked Data?

- Tasks and Tools

4. How to consume Linked Data?

- Tasks and Tools

Linked Data Principles

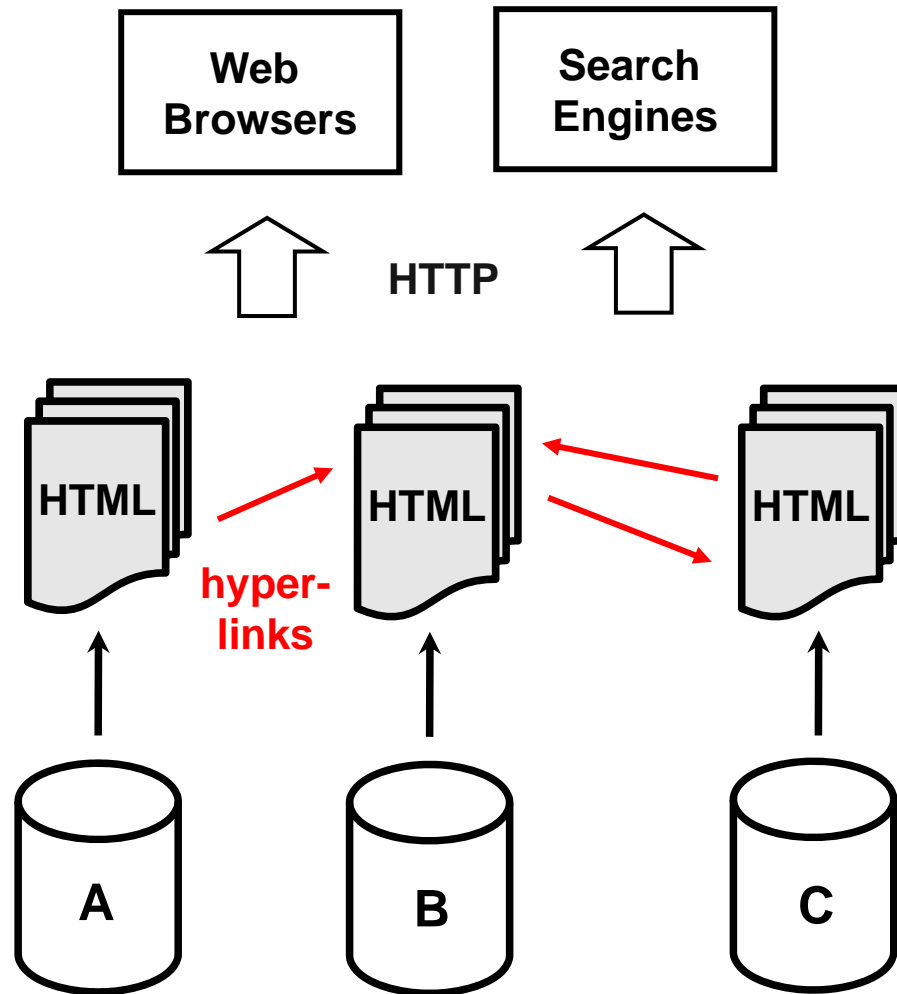
Set of best practices for publishing structured data on the Web in accordance with the general architecture of the Web.



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, <http://www.w3.org/DesignIssues/LinkedData.html>, 2006

Architecture of the classic Web



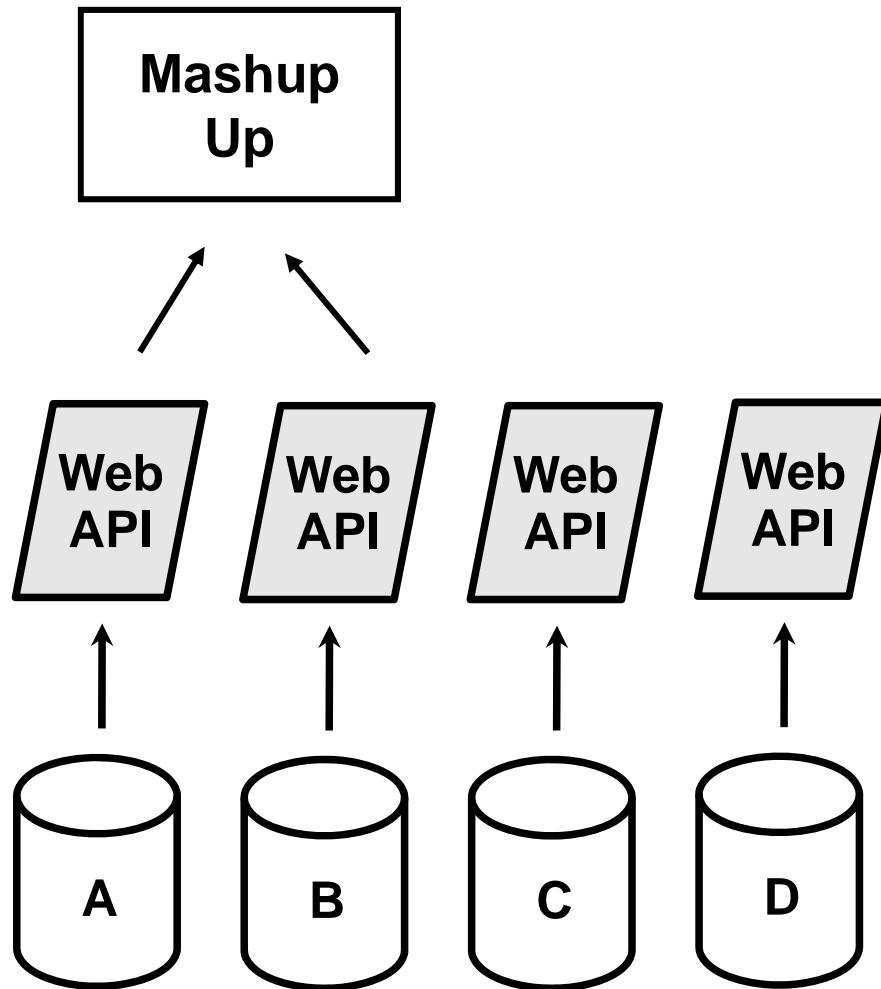
Single global information space

Small set of simple standards

- 1. HTML as document format**
- 2. HTTP URLs as**
 - globally unique IDs
 - retrieval mechanism
- 3. Hyperlinks to connect everything**



Web 2.0 APIs and Mashups

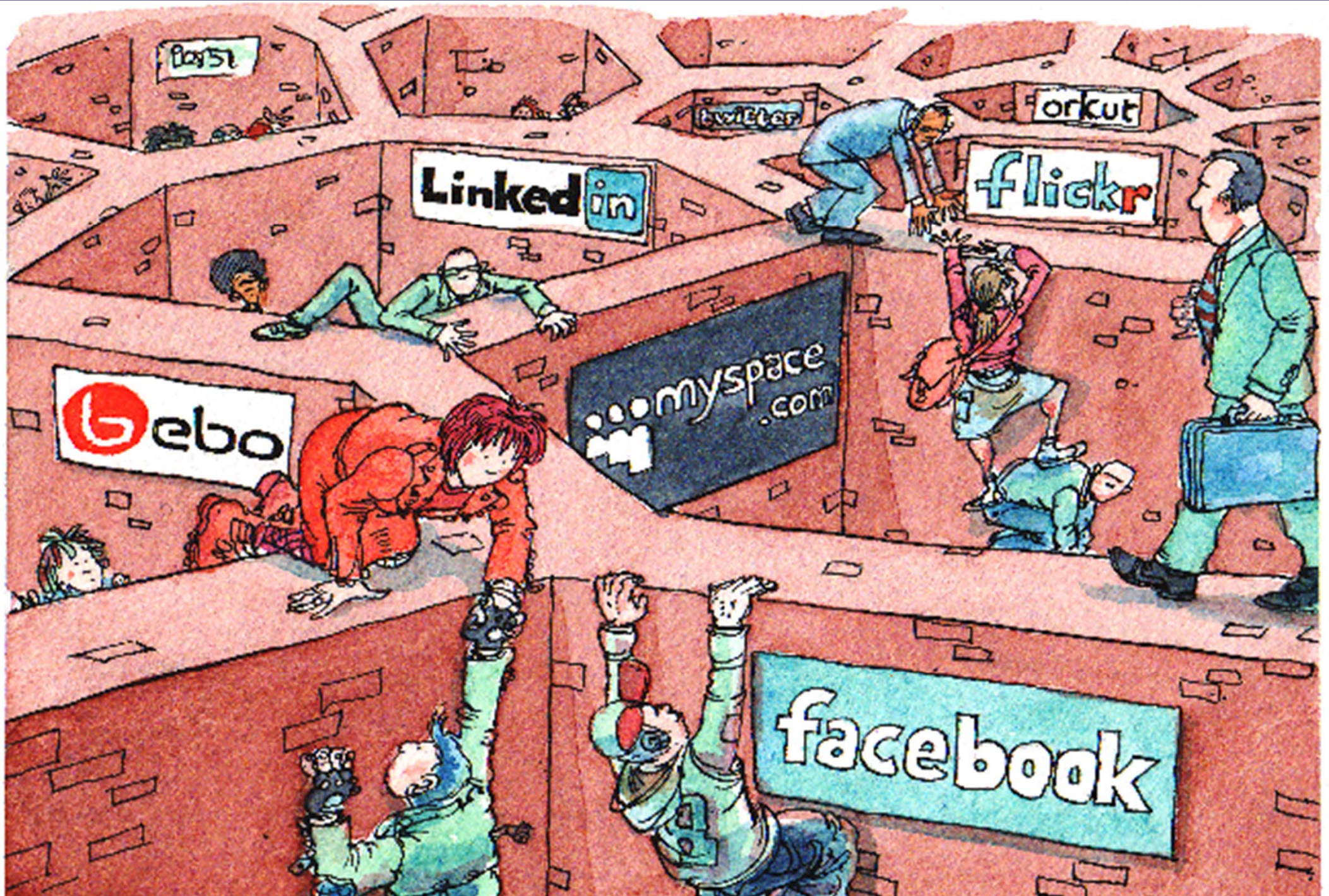


No single global dataspace

Shortcomings

1. APIs have proprietary interfaces
2. Mashups are based on a fixed set of data sources
3. No hyperlinks between data items within different APIs

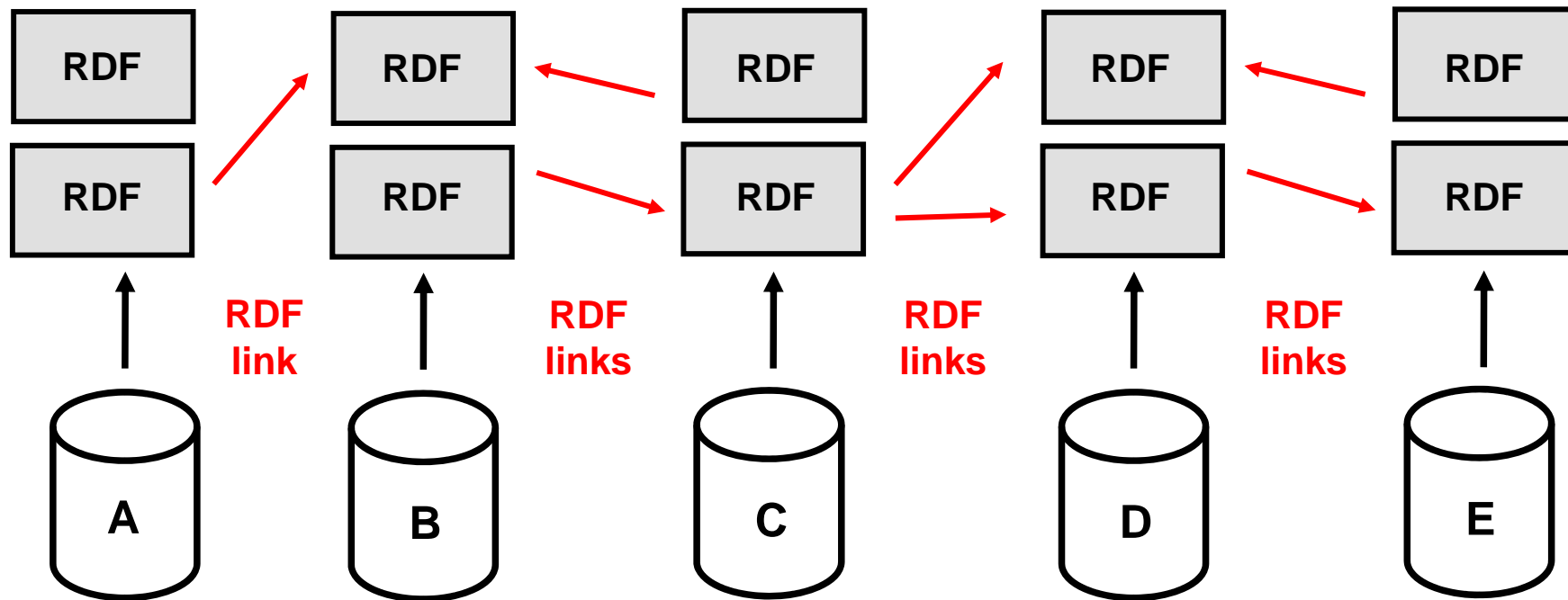
Web APIs slice the Web into Walled Gardens



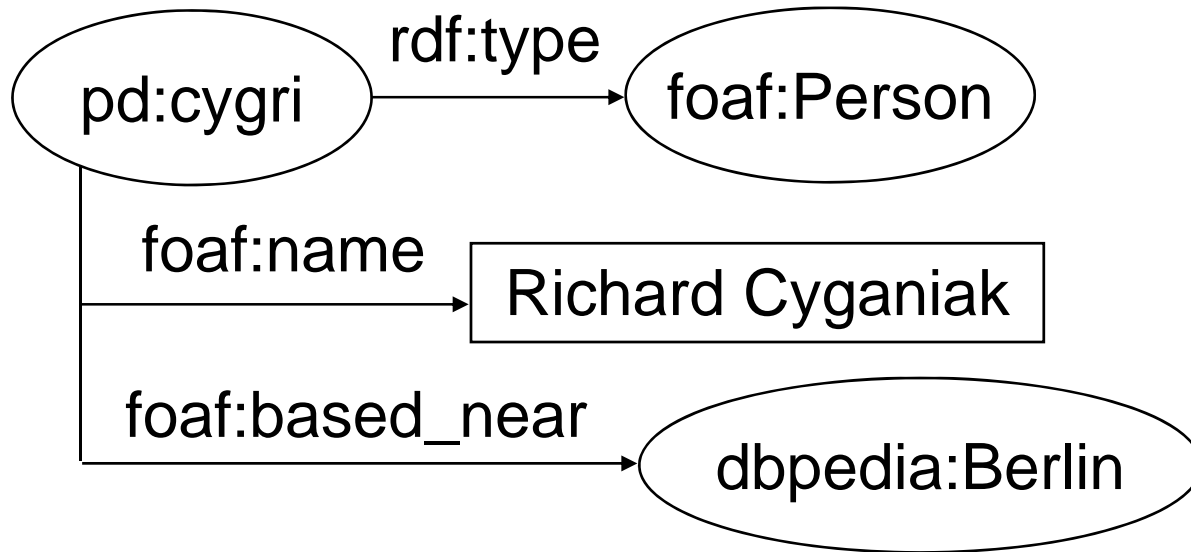


Extend the Web with a **single global dataspace**

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.

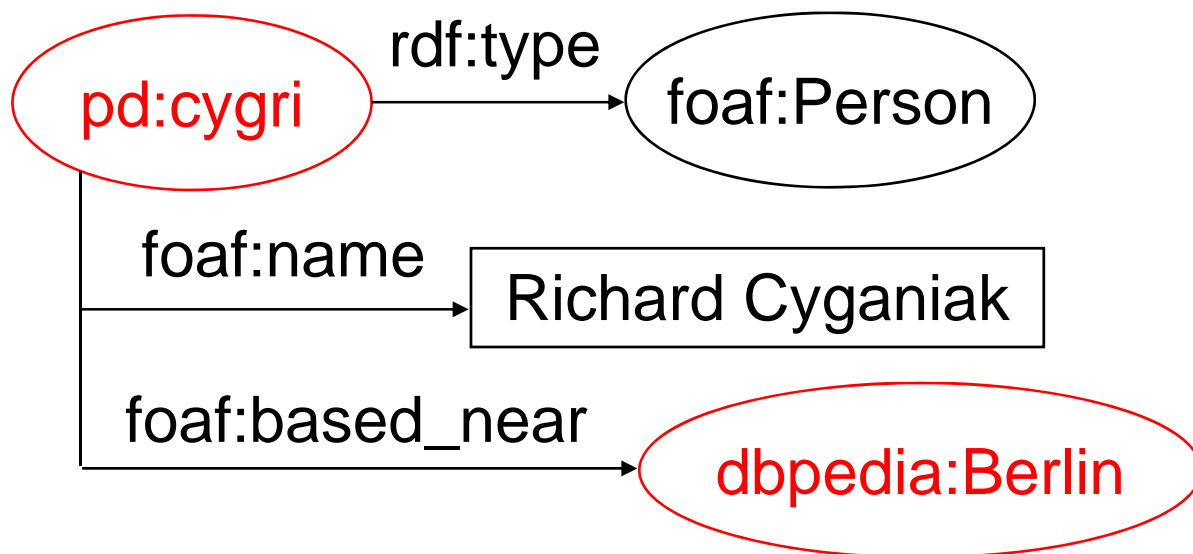


The Basis: RDF Data Model



Flexible graph-based data model.

Data items are identified with HTTP URIs

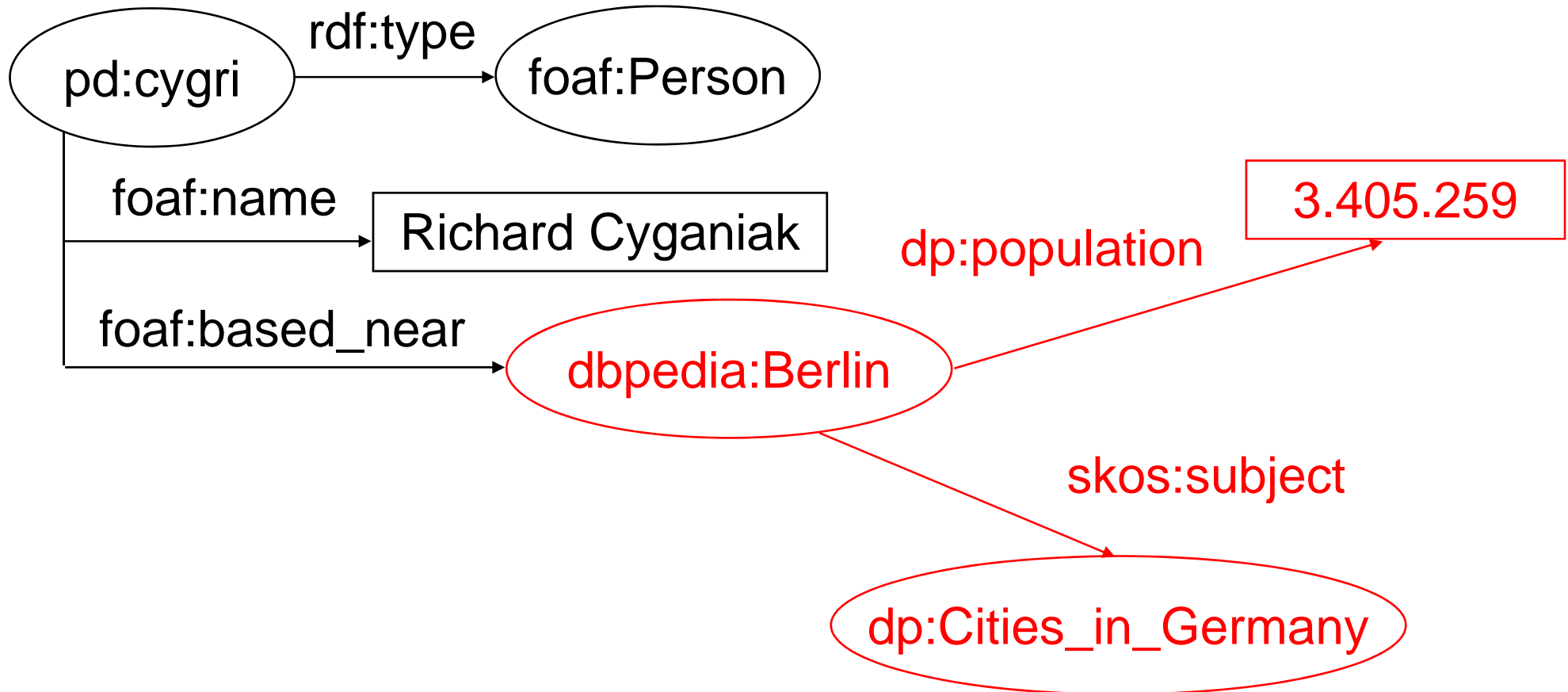


HTTP URIs take the role of global primary keys.

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

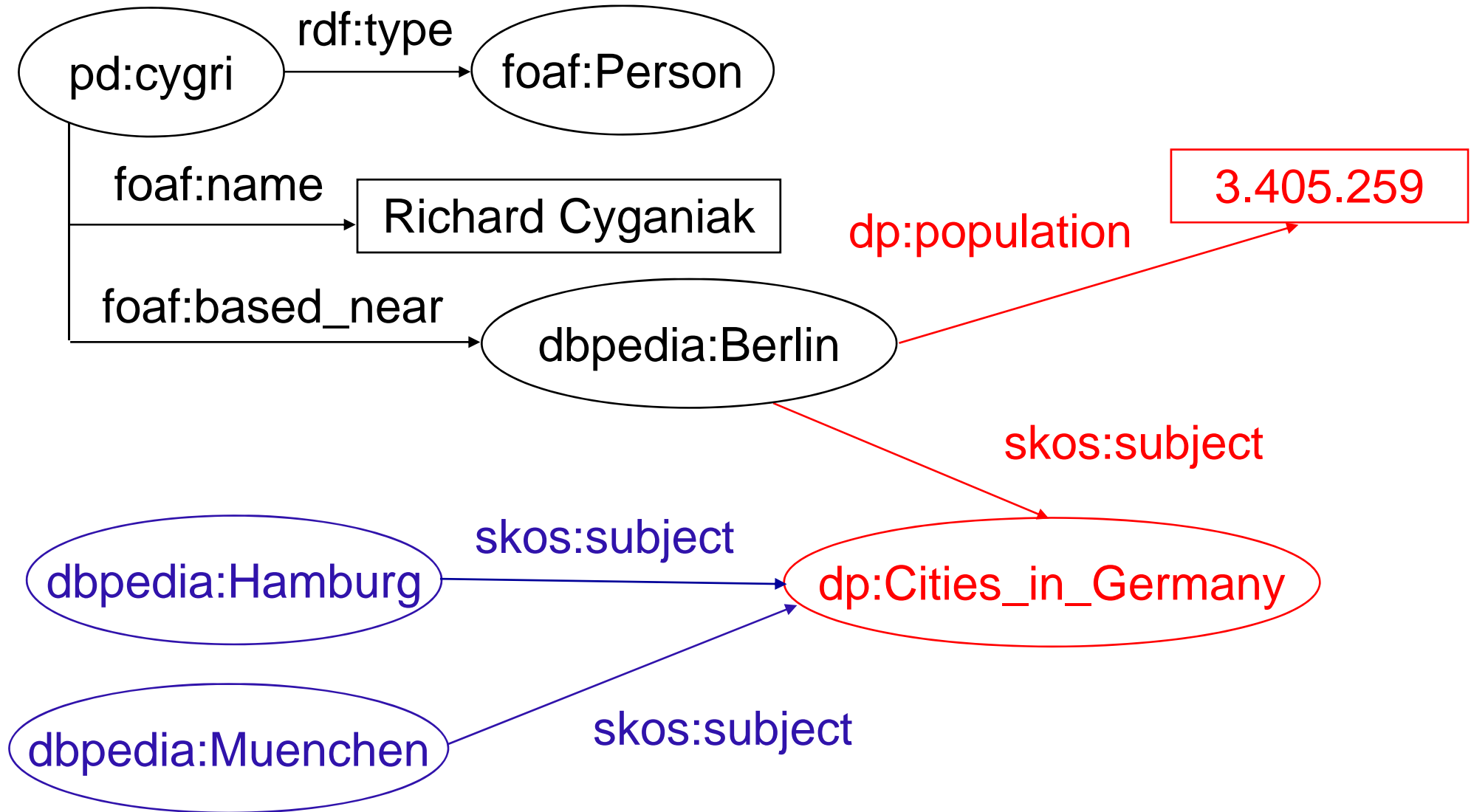
dbpedia:Berlin = <http://dbpedia.org/resource/Berlin>

Resolving URIs over the Web




The HTTP protocol brings together identification and retrieval again.

Following Links deeper into the Web



Richard Cyganiak

URI:

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

Berlin

URI:

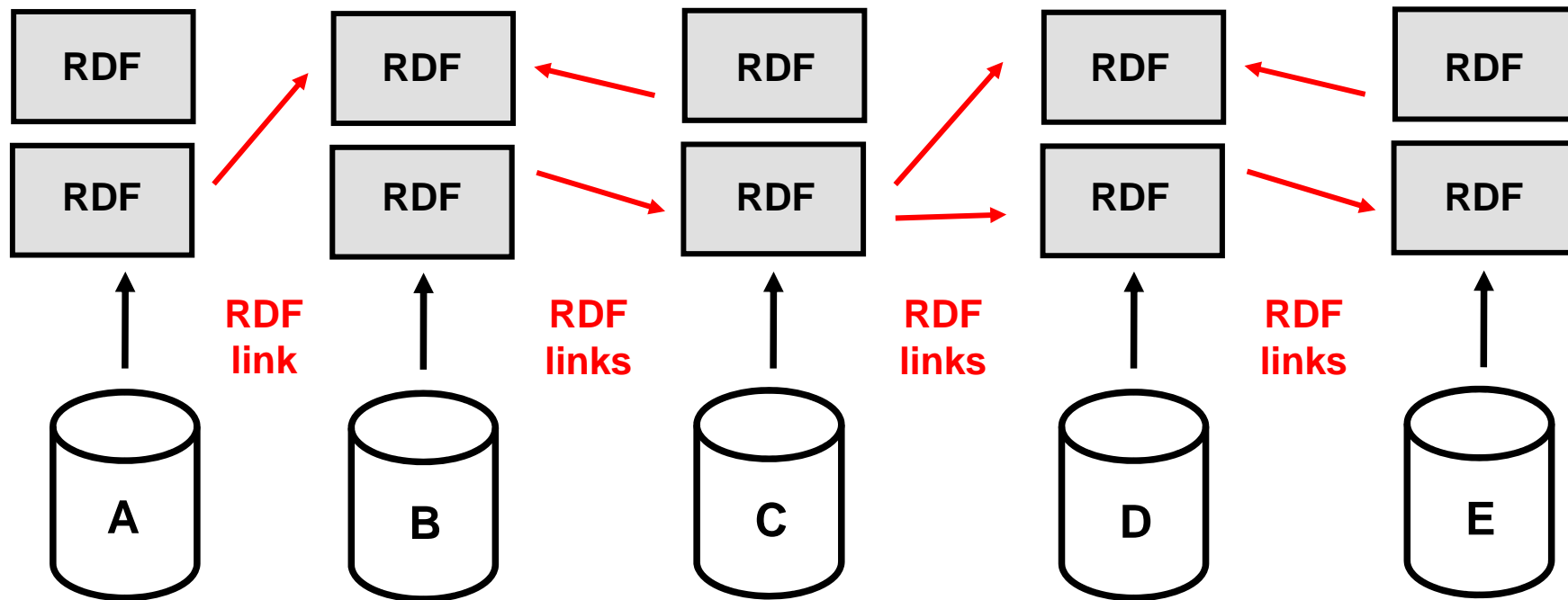
Property	Value	Sources
population	3398888	G2
type	http://dbpedia.org/City	G2
comment	Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union.	G2
comment	Berlin ist die 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

- **Global, distributed dataspace build on a simple set of standards**
 - RDF, URIs, HTTP
- **Entities are connected by links**
 - creating a global data graph that spans data sources and
 - enables the discovery of new data sources
- **Provides for data-coexistence**
 - Everyone can publish data to the Web of Linked Data
 - Everyone can express their personal view on things
 - Everybody can use the vocabularies/schema that they like

Linked Data Deployment on the Web

■ Is this real?



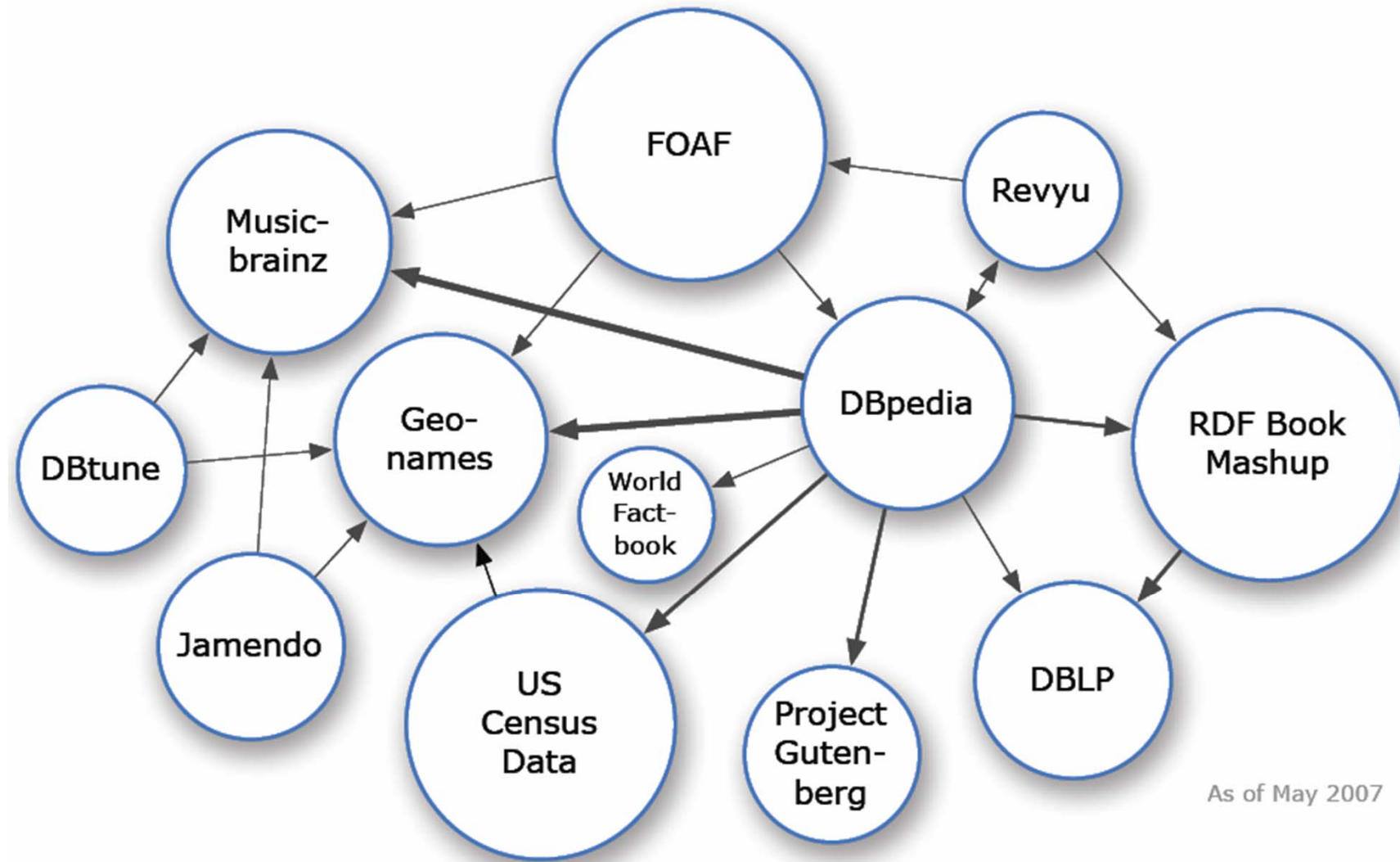
W3C Linking Open Data Project



■ **Grassroots community effort to**

- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources

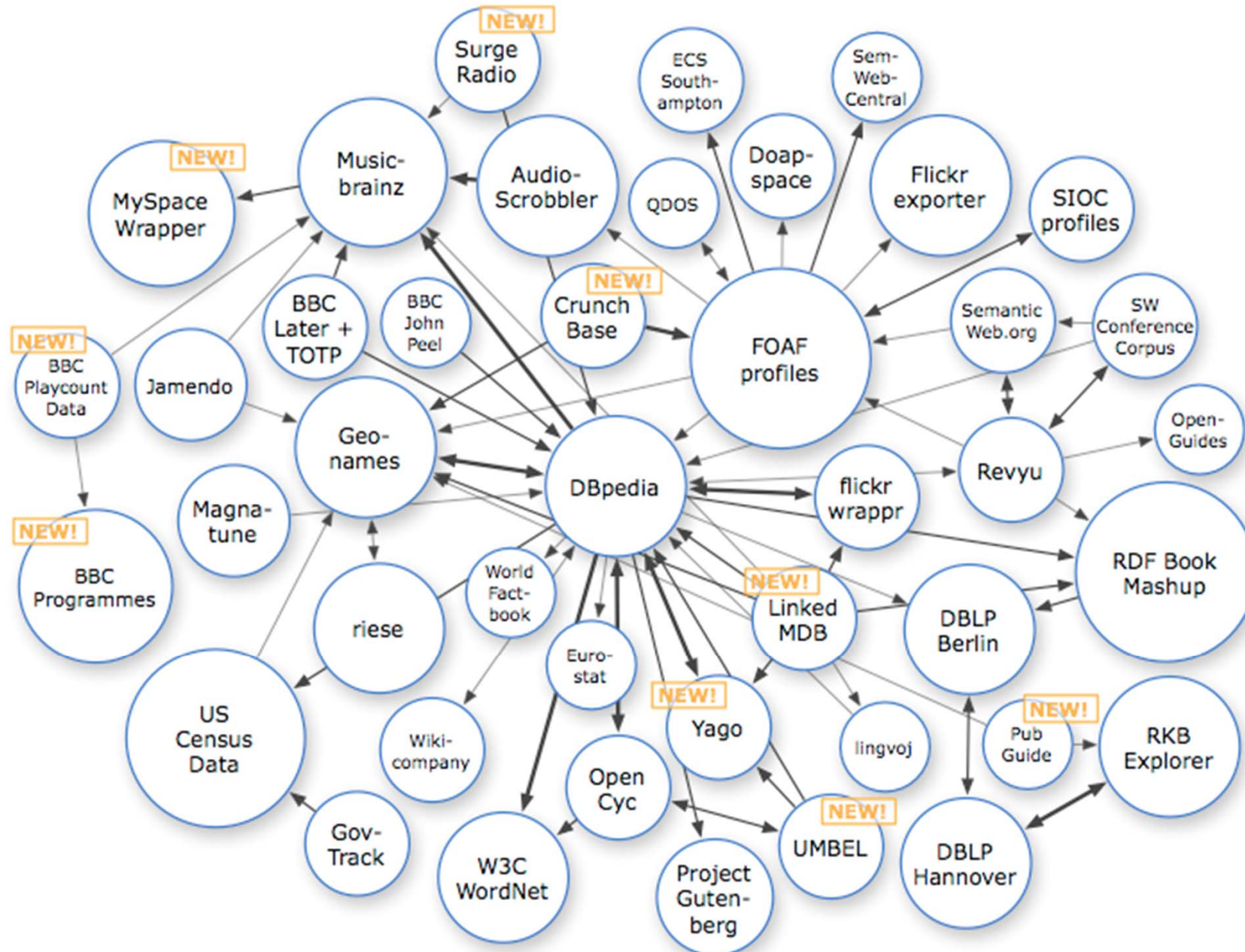
LOD Datasets on the Web: May 2007



As of May 2007

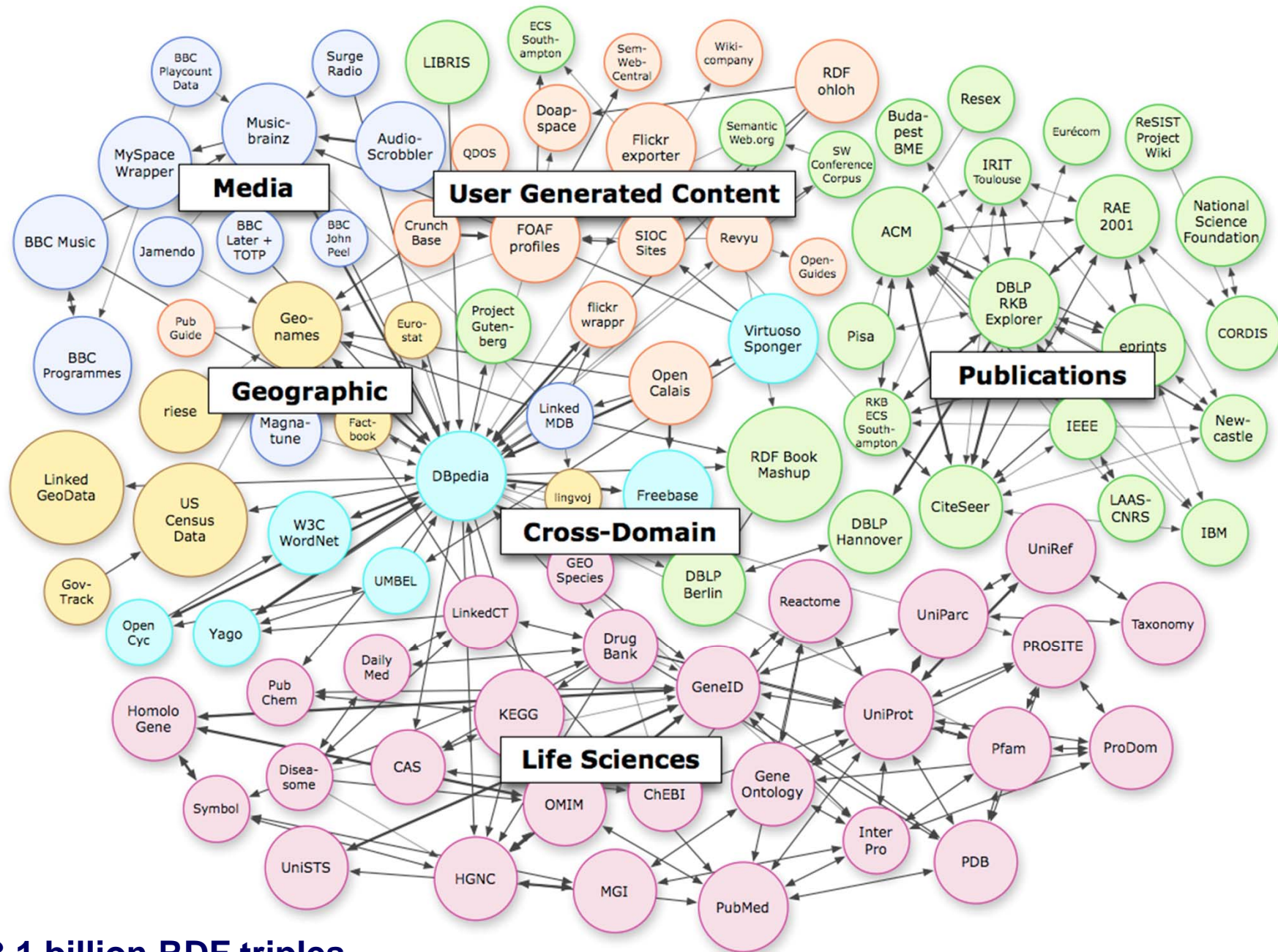
- Over 500 million RDF triples
- Around 120,000 RDF links between data sources

LOD Datasets on the Web: September 2008



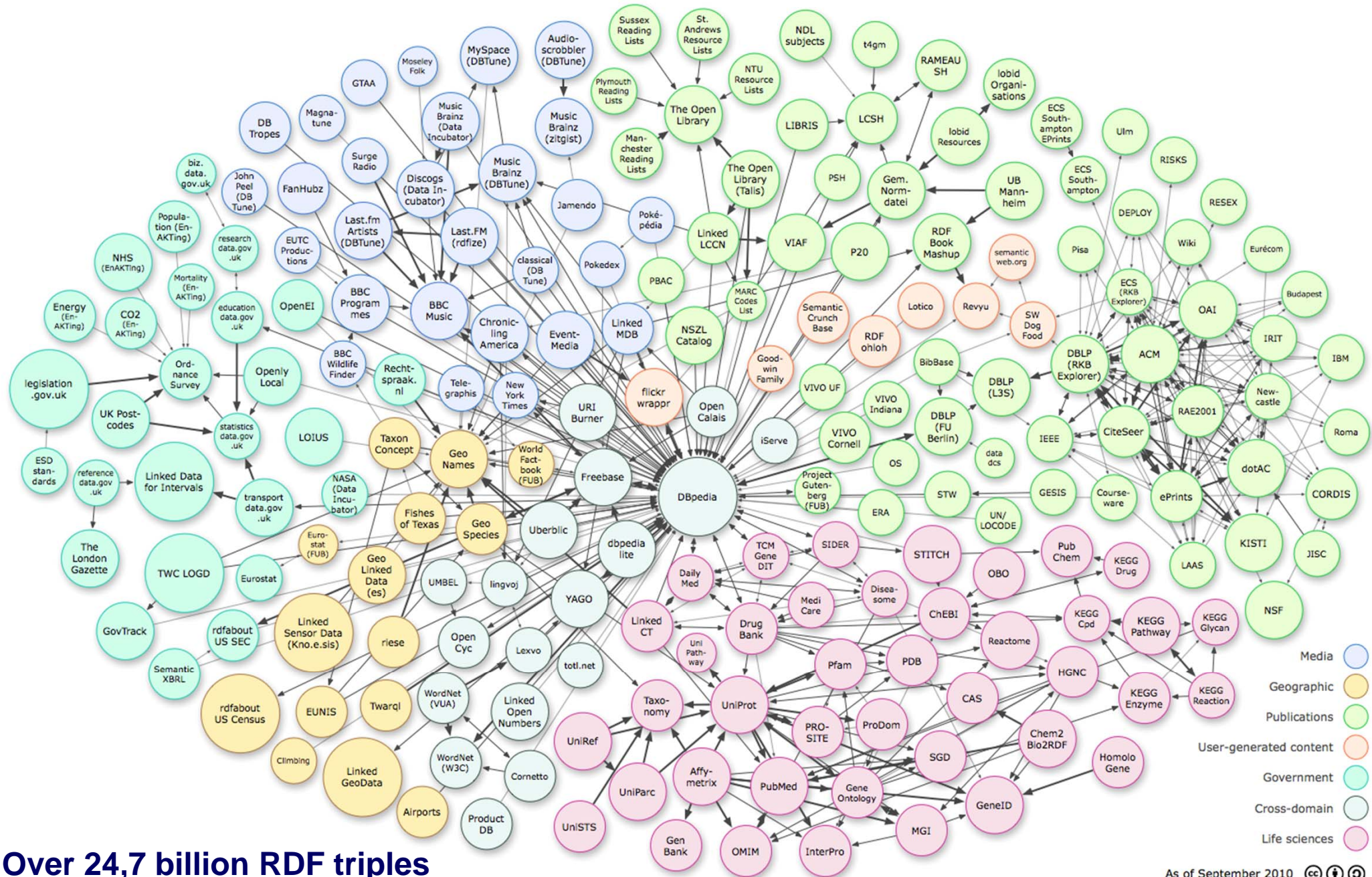
As of September 2008

LOD Datasets on the Web: July 2009



- Over 13.1 billion RDF triples
- Over 142 million RDF links between data sources

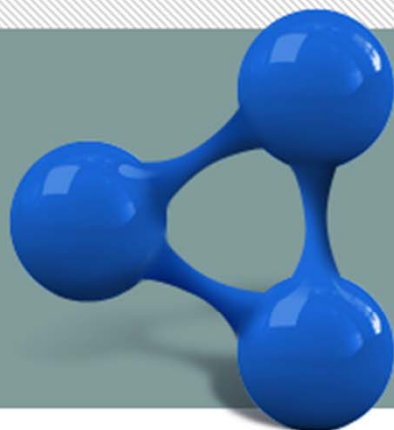
LOD Datasets on the Web: September 2010



- Over 24,7 billion RDF triples
- Over 436 million RDF links between data sources

Unlocking innovation

Working with UK Public
Sector information and data



Advised by Sir Tim Berners-Lee and Professor Nigel Shadbolt and others, government is opening up data for reuse. This site seeks to give a way into the wealth of government data and is under constant development. We want to work with you to make it better.

We're very aware that there are more people like you outside of government who have the skills and abilities to make wonderful things out of public data. These are our first steps in building a collaborative relationship with you.

Latest news:

- Read about our latest [site changes](#)
- find out how the [data.gov.uk team has been getting involved with the community](#)
- listen to a [Podcast on setting up data.gov.uk](#)

Search Data

Enter keyword(s)

Search

e.g. education, NHS, crime, transport, environment

Powered by: [CKAN](#)

Browse for Data

[List all datasets](#)

[By Public Body](#)

[Common tags](#)

Subscribe by [RSS](#)



Community
[Log in / Sign up](#)

[Local Data Panel](#)



What is the Semantic Web?

Combining different data sources has never been easy but the Semantic Web will enable data to be joined easily across boundaries.

[Read more](#)

Digital Engagement
[Twitter stream](#)

LINKING OPEN GOVERNMENT DATA

VIEW MORE ▶



Most Popular Datasets

1. Worldwide M1+ Earthquakes, Past 7 Days
2. U.S. Overseas Loans and Grants (Greenbook)
3. Latest Volumes of Foreign Relations of the...
4. OSHA Data Initiative - Establishment...
5. IT Dashboard - Federal IT Spending (major...

SEARCH OUR CATALOGS

Search our catalogs..

SEARCH ▶

APPS



With so much government data to work with, developers are creating a wide variety of applications, mashups, and visualizations. From crime statistics by neighborhood to the best

COMMUNITY

Data.gov is leading the way in democratizing public sector data and driving innovation. The data is being surfaced from many locations making the Government data stores available to researchers to perform their own analysis. Developers are finding good uses for the datasets, providing interesting and useful applications that allow for new views and public analysis. This is a work in progress, but this movement is spreading to cities, states, and other countries. After just one year a community is born around open government data.

Just look at the numbers:

SEMANTIC WEB

As the Web of linked documents evolves to include the Web of linked data, we're working to maximize the potential of Semantic Web technologies to realize the promise of Linked Open Government Data.

Thanks to our collaboration with the **Tetherless World Constellation** at the **Rensselaer Polytechnic Institute**, Data.gov is now hosting



Uptake in the Libraries Community

■ Institutions publishing Linked Data

- Library of Congress (subject headings)
- Schwedische Nationalbibliothek (Libris - catalog)
- Hungarian National Library (OPAC and Digital Library)
- Deutschen Zentralbibliothek für Wirtschaftswissenschaften (subject headings)
- Deutsche Nationalbibliothek (PND dataset and subject headings)
- Europeana project is moving towards Linked Data

■ W3C Library Linked Data Incubator Group

■ Open Archives ORE Standard

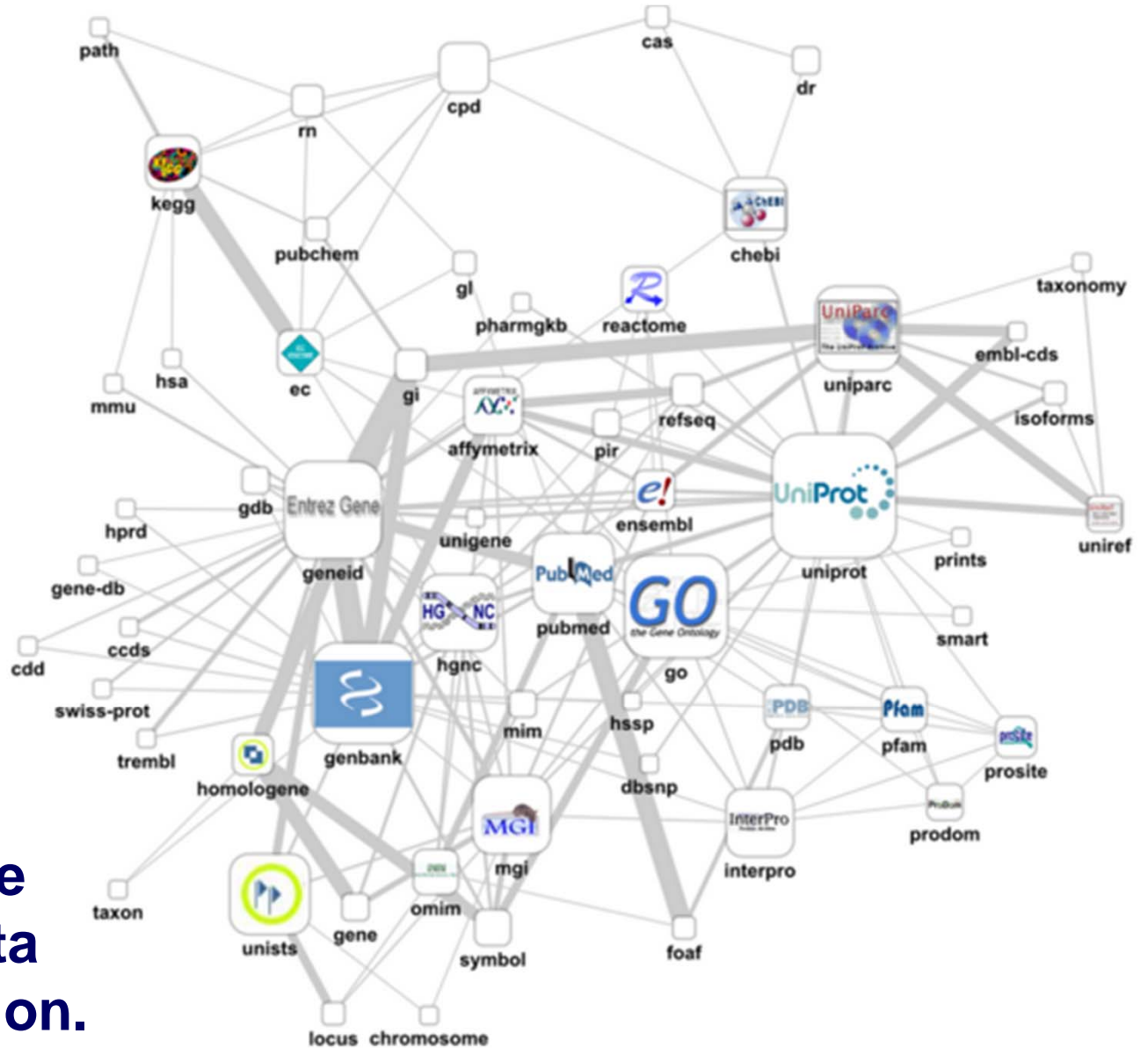
■ Goals:

1. Integrate Library Catalogs on global scale.
2. Interconnect resources between repositories
(by topic, by location, by historical period, by ...).

Uptake in Life Sciences

- W3C Linking Open Drug Data Effort
- Bio2RDF Project
- Allen Brain Atlas

- **Goal:** Smoothly integrate internal and external data in a pay-as-you-go-fashion.



Uptake in the Media Industry

The New York Times

Linked Open Data BETA

Search data.nytimes.com

data.nytimes.com

For the last 150 years, The New York Times has maintained one of the most authoritative news vocabularies ever developed. In 2009, we began to publish this vocabulary as linked open data.

The Data

As of 13 January 2010, The New York Times has published approximately 10,000 subject headings as linked open data under a CC BY license. We provide both RDF documents and a human-friendly HTML versions. The table below gives a breakdown of the various tag types and mapping strategies on data.nytimes.com.

Type	Manually Mapped Tags	Automatically Mapped Tags
People	4,978	0
Organizations	1,489	1,592
Locations	1,910	0

- Publish data as RDF/XML or RDFa
- **Goal:** Drive traffic to websites via search engines

BBC Text only Help Search Explore the BBC Sign in Register

MUSIC BETA GENRES ARTISTS REVIEWS NEWS BLOG QUICK FIND Enter an artist name ...

BBC TWO
Later Live
 With Mumford & Sons, Hole, Joanna Newsom, Ian Hunter and Angelique Kidjo, plus a chat with Iggy Pop.

MORE MUSIC ON THE BBC 1-3 OF 20

- LATER LIVE
- GORILLAZ - LIVE IN THE FLESH ON LATER...
- KATIE MELUA - WEEKEND WOGAN

BROWSE BY GENRE

- Classic Pop & Rock
- Classical
- Country
- Dance & Electronica
- Desi
- Easy Listening, Soundtracks & Musicals
- Folk
- Hip Hop, RnB & Dancehall
- Jazz & Blues
- Pop & Chart
- Rock & Indie
- Soul & Reggae
- World

Most Played Artists On The BBC

3-9 MAY 2010

1 Tinie Tempah 2 Gorillaz 3 Roll Deep 4 Rihanna 5 Kelis 6 Biffy Clyro 7 Chipmunk 8 Dizzee Rascal 9

Information displayed about artists played on BBC programmes is incomplete at present. Find out more about this artist play count information.

North Korea's Exit Strategy | Facebook: Friend or Foe? | BP Oil Spill Timeline | Subscribe | msnbc.com

Newsweek
 What matters most

STORIES TOPICS AUTHORS SEARCH

FedEx
 Ground

Faster to more locations than UPS Ground.®

UPS is a registered trademark of United Parcel Service

SAVE NOW ▶

Week NOW Subscribe now and save up to 78%

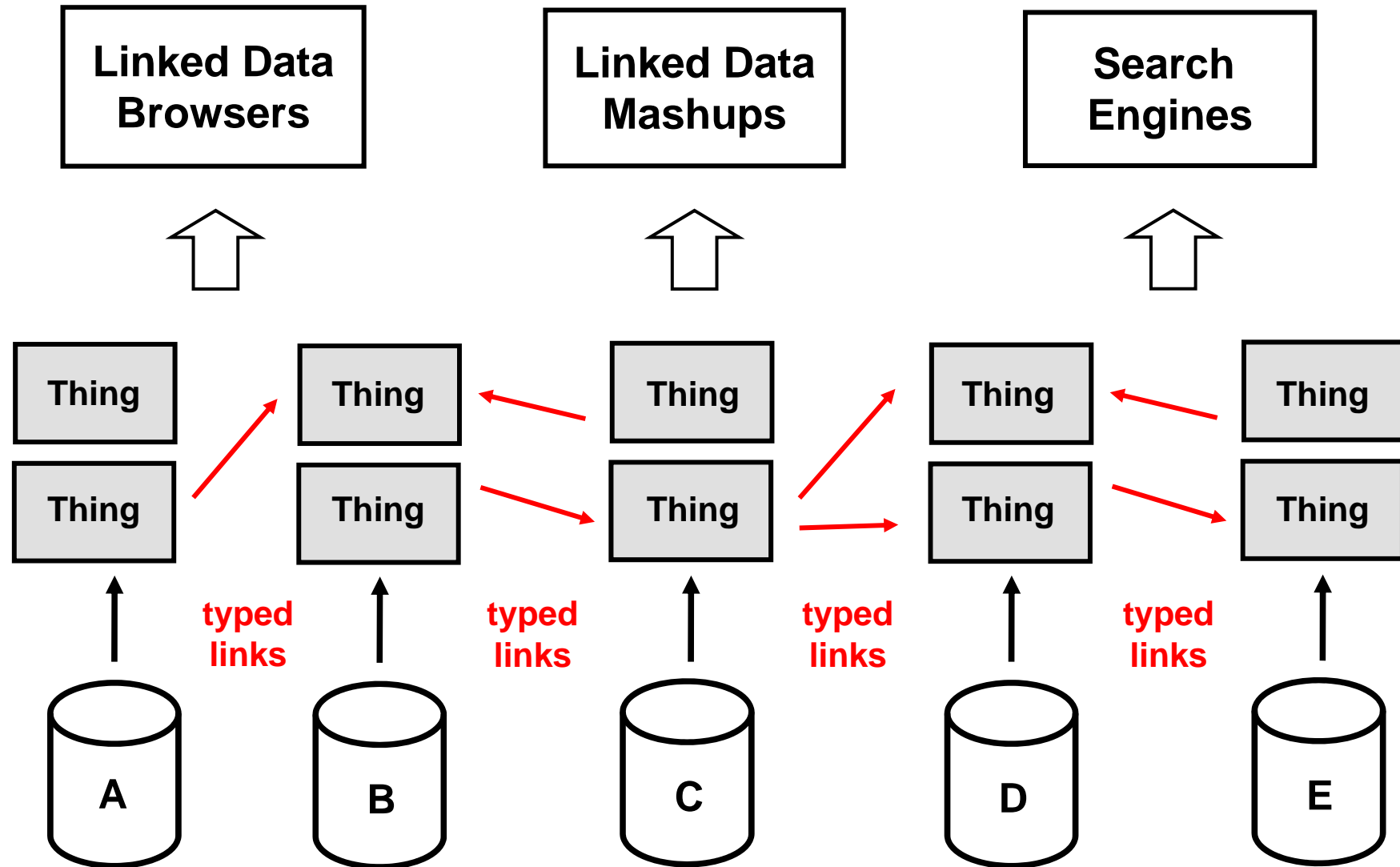
are. We're Queer. We're Retiring.

To Meet The Demands Of America's Aging Gay Population, Developers Are Now Targeting The Lgbt Market With Everything From Active Adult Rental Apartments To Retirement Communities That Promise Life-Long Care. **More >**

Business
 by Linda Stern
 May 26, 2010
 43

Linked Data Applications

■ What can I do with this?



Linked Data Browsers

Provide for navigating between data sources and for exploring the dataspace.

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

Tim Berners-Lee

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

- [Person](#)
- <http://www.w3.org/2000/10/swap/pim/contact#Male>

[label](#)

- [Tim Berners-Lee](#)

[sameAs](#)

- [Tim Berners-Lee \(also at www4.wiwiw.fu-berlin.de\)](#)

[image](#)



[Weblinks](#)

- <http://www.w3.org/People/Berners-Lee/>

[name](#)

- [Tim Berners-Lee](#)
- [Timothy Berners-Lee](#)
- [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](#)
- [timbl](#)

[personal mailbox](#)

- <mailto:timbl@w3.org>

[seeAlso](#)

- [Tim Berners-Lee's FOAF file](#)
- [Tim Berners-Lee's FOAF file](#)

[is seeAlso of](#)

- [Tim Berners-Lee](#)

Web of Data Search Engines

Crawl the dataspace and provide best-effort query answers over crawled data.

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

[Add More Info](#)
[Start New](#)
[Options](#)
[Order](#)
[Permalink](#)

Chris Bizer

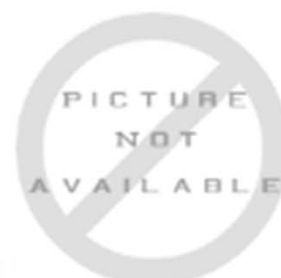
picture:



[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 Relational 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]

[NG4J: Named Graphs API for Java](#) [16]

1 [Chris Bizer - Free Uni](#)

http://videlectures.net/chris_

2 [Chris Bizer - semantic](#)

<http://ontoworld.org/wiki/Chris>

3 [Untitled document](#) 6 f

[BOSS http://www.facebook](http://www.facebook)

4 [Chris Bizer - semantic](#)

<http://semanticweb.org/wiki/Ch>

5 [Chris Bizer - LinkedIn](#)

[BOSS http://www.linkedin](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/ph](http://flickr.com/ph)

9 [Untitled document](#) 8 f

<http://data.semanticweb.org/cc>

10 [Chris Bizer](#) 6 facts | 20

[BOSS http://ebiquity.umbc](http://ebiquity.umbc)

<-

1

2

->

<http://example.loc/doc>

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](#)

Whatfettle marras?

3. How to publish Linked Data?

Tasks:

1. **Make data available as RDF via HTTP**
2. **Set RDF links pointing at other data sources**
3. **Make your data self-descriptive**

■ **How to publish Linked Data Tutorial**

<http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/>

3.1 Make Data available as RDF via HTTP

Ready to use tools (examples)

1. D2R Server

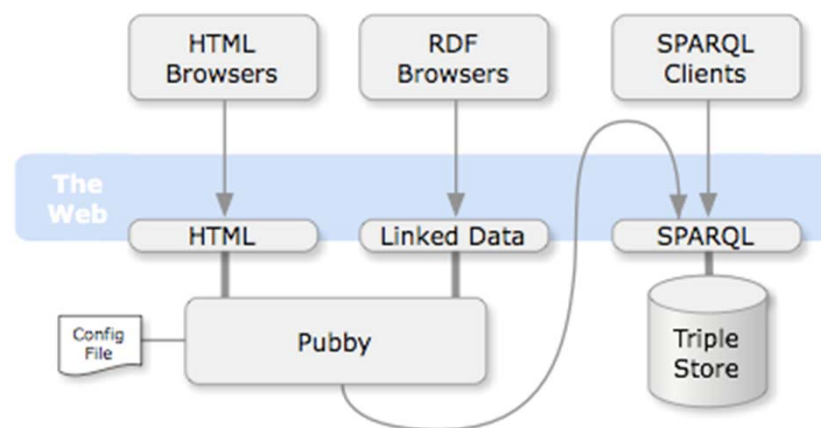
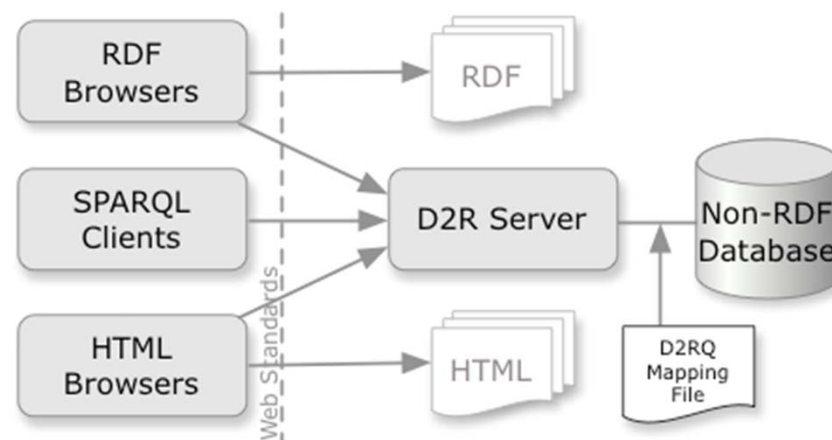
- provides for mapping relational databases into RDF and for serving them as Linked Data

2. Pubby

- Linked Data Frontend for SPARQL Endpoints

3. More tools

- <http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/PublishingTools>



3.2 Set RDF links pointing at other data sources

■ Examples of RDF links

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

```
<http://richard.cyganiak.de/foaf.rdf#cygri> foaf:topic_interest  
<http://dbpedia.org/resource/Semantic_Web> .
```

```
<http://example-bookshop.com/book006251587X> owl:sameAs  
<http://www4.wiwiss.fu-berlin.de/bookmashup/books/006251587X> .
```


How to generate RDF links?

■ Pattern-based Approaches

- Exploit naming conventions within URIs (for instance ISBNs, ISINs, ...)

■ Similarity-based Approaches

- Compare items within different data sources using various similarity metrics

Ready to use tools (Examples)

1. Silk – Link Discovery Framework

- provides a declarative language for specifying link conditions which may combine different similarity metrics
- Silk Single Machine, Silk MapReduce

2. More tools

- <http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining>

3.3 Make your Data Self-Descriptive

■ Increase the usefulness of your data and ease data integration

■ Aspects of self-descriptiveness

1. Enable clients to retrieve the schema
2. Reuse terms from common vocabularies
3. Publish schema mappings for proprietary terms
4. Provide provenance metadata
5. Provide licensing metadata
6. Provide data-set-level metadata using void
7. Refer to additional access methods using void

■ Statistics about the compliance with these best practices

- <http://lod-cloud.net/state/>

Enable Clients to retrieve the Schema

Clients can resolve the URIs that identify vocabulary terms in order to get their RDFS or OWL definitions.

Some data on the Web

```
<http://richard.cyganiak.de/foaf.rdf#cygri>  
  foaf:name "Richard Cyganiak" ;  
  rdf:type <http://xmlns.com/foaf/0.1/Person> .
```



Resolve unknown term

`http://xmlns.com/foaf/0.1/Person`

RDFS or OWL definition

```
<http://xmlns.com/foaf/0.1/Person>  
  rdf:type owl:Class ;  
  rdfs:label "Person";  
  rdfs:subClassOf <http://xmlns.com/foaf/0.1/Agent> ;  
  rdfs:subClassOf <http://xmlns.com/wordnet/1.6/Agent> .
```

Reuse Terms from Common Vocabularies

■ Common Vocabularies

- **Friend-of-a-Friend** for describing people and their social network
- **SIOC** for describing forums and blogs
- **SKOS** for representing topic taxonomies
- **Organization Ontology** for describing the structure of organizations
- **GoodRelations** provides terms for describing products and business entities
- **Music Ontology** for describing artists, albums, and performances
- **Review Vocabulary** provides terms for representing reviews

■ Common sources of identifiers (URIs) for real world objects

- **LinkedGeoData** and **Geonames** locations
- **GeneID** and **UniProt** life science identifiers
- **DBpedia** wide range of things

Publish Schema Mappings on the Web

```
<http://xmlns.com/foaf/0.1/Person>  
owl:equivalentClass  
<http://dbpedia.org/ontology/Person> .
```

■ Simple Mappings: OWL, RDFS, SKOS

- owl:equivalentClass, owl:equivalentProperty,
- rdfs:subClassOf, rdfs:subPropertyOf
- skos:broadMatch, skos:narrowMatch

■ Complex Mappings: R2R

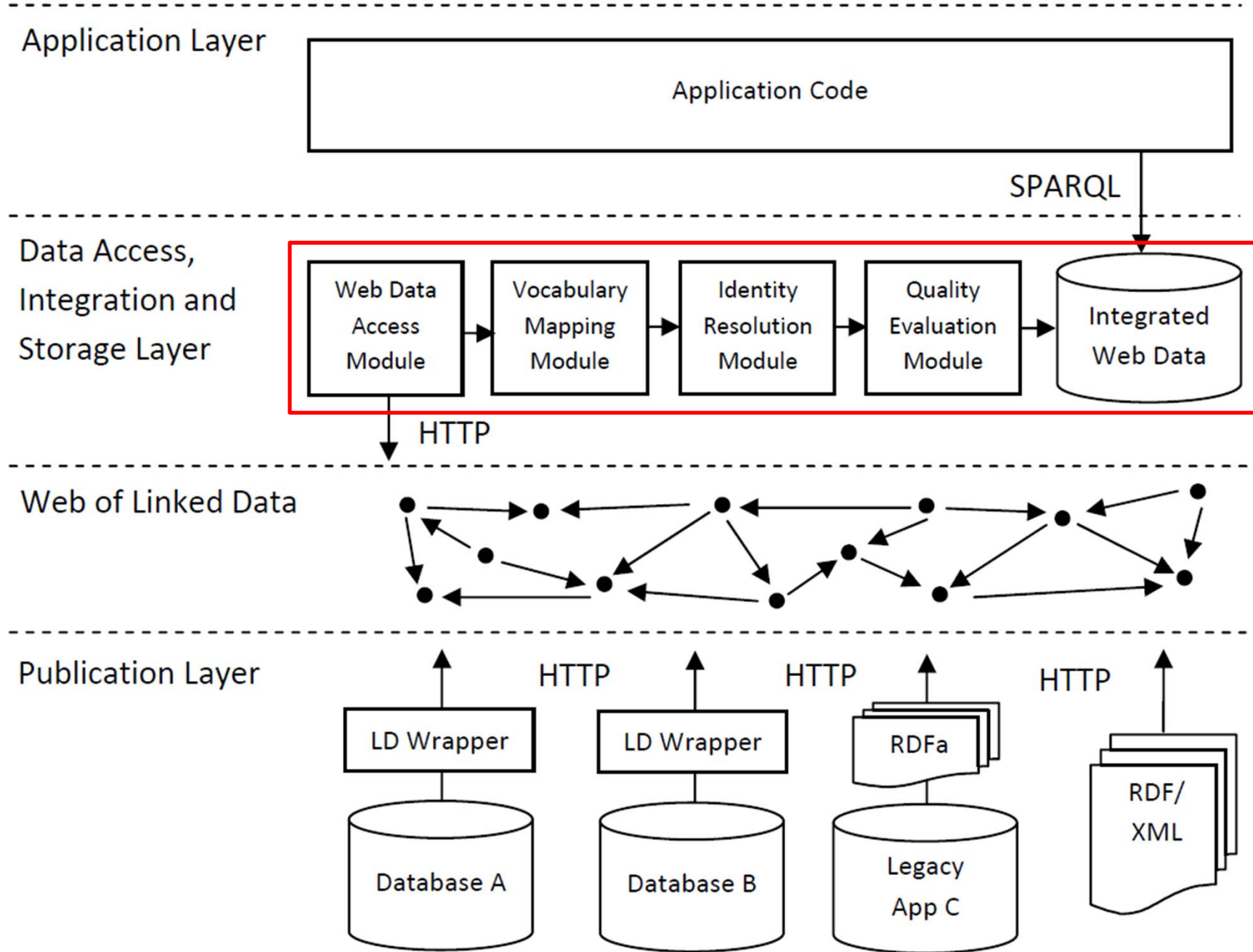
- provides value transformation functions
- structural transformations

4. How to consume Linked Data?

Tomorrow:

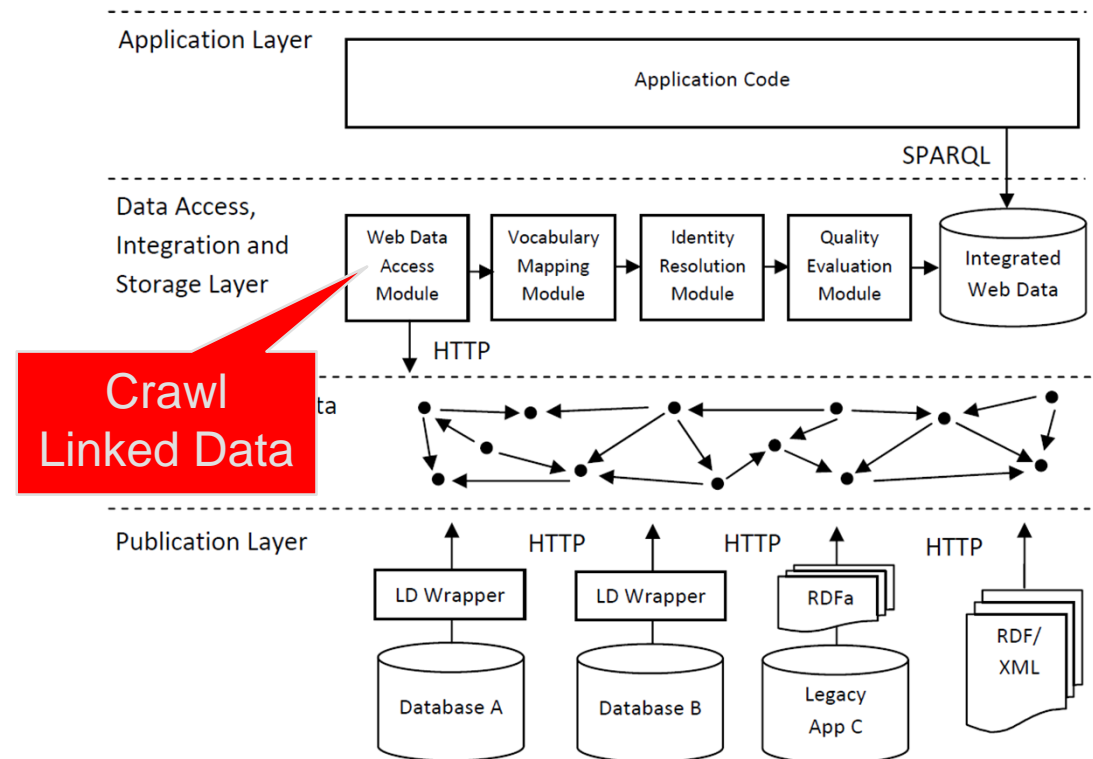
**First International Workshop on
Consuming Linked Data (COLD 2010)**

Task involved in Linked Data Consumption



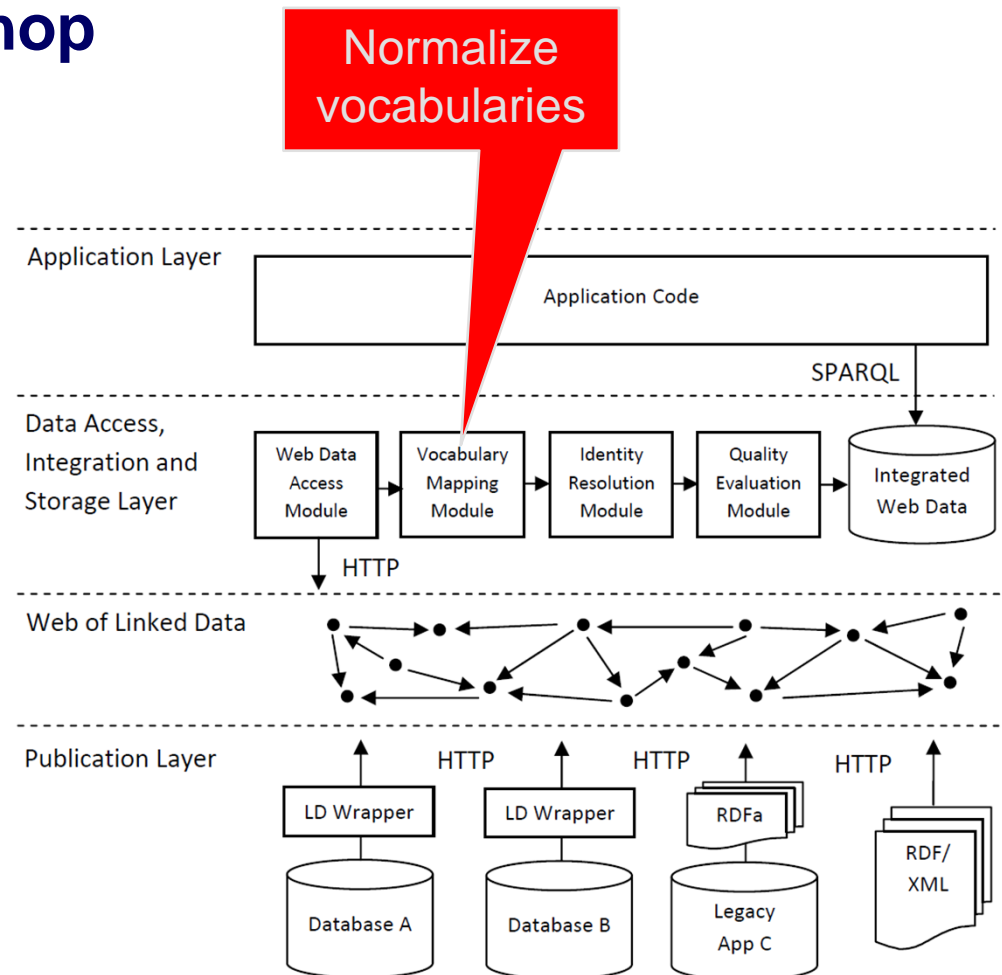
LDspider

- Flexible open-source Linked Data crawler
- Crawls RDF/XML and RDFa
- Poster about LDspider at ISWC poster session



R2R Framework

- Tool for translating RDF data between different vocabularies
- Provides for publication and discovery of mappings on the Web
- Talk about R2R at COLD workshop



The Dataspace Vision

Alternative to classic data integration systems in order to cope with growing number of data sources.

■ Properties of dataspaces

- may contain any kind of data (structured, semi-structured, unstructured)
- require no upfront investment into a global schema
- provide for data-coexistence
- give best effort answers to queries
- rely on pay-as-you-go data integration

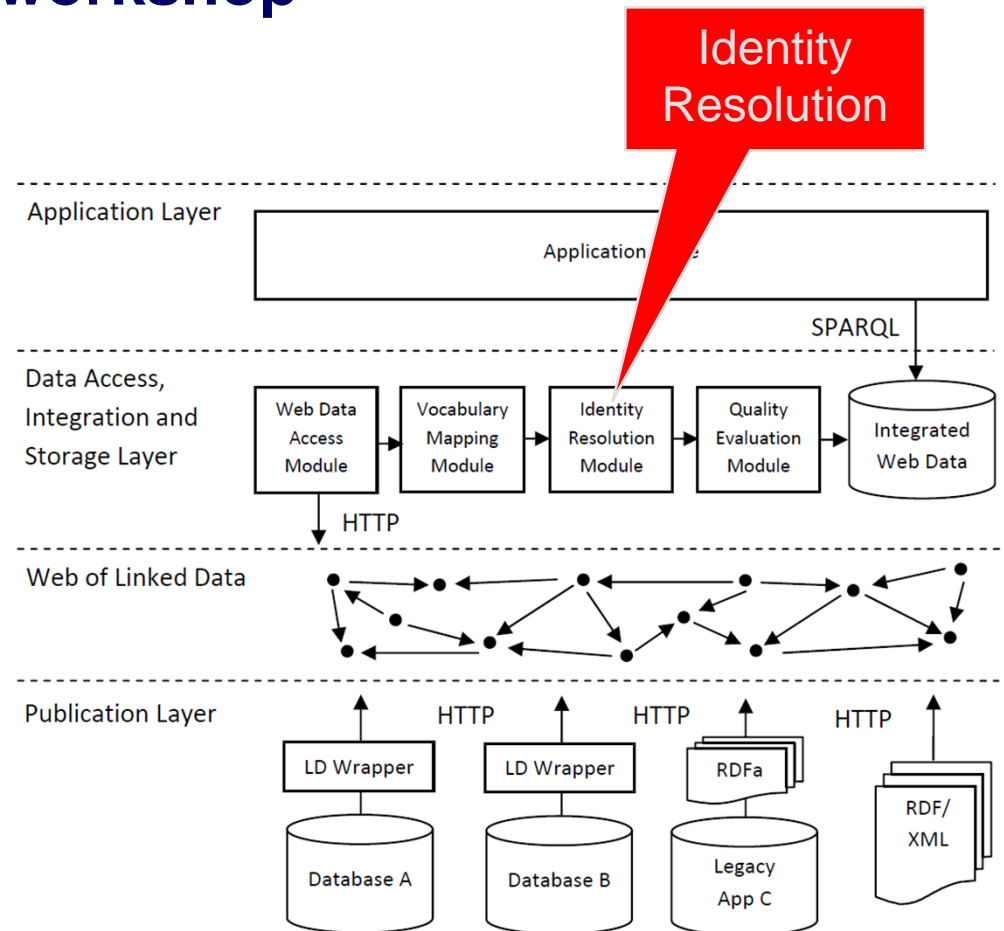
Franklin, M., Halevy, A., and Maier, D.: From Databases to Dataspaces A new Abstraction for Information Management, SIGMOD Rec. 2005.

Madhavan, J., et al.: Web-scale Data Integration: You Can Only Afford to Pay As You Go, CIDR 2007



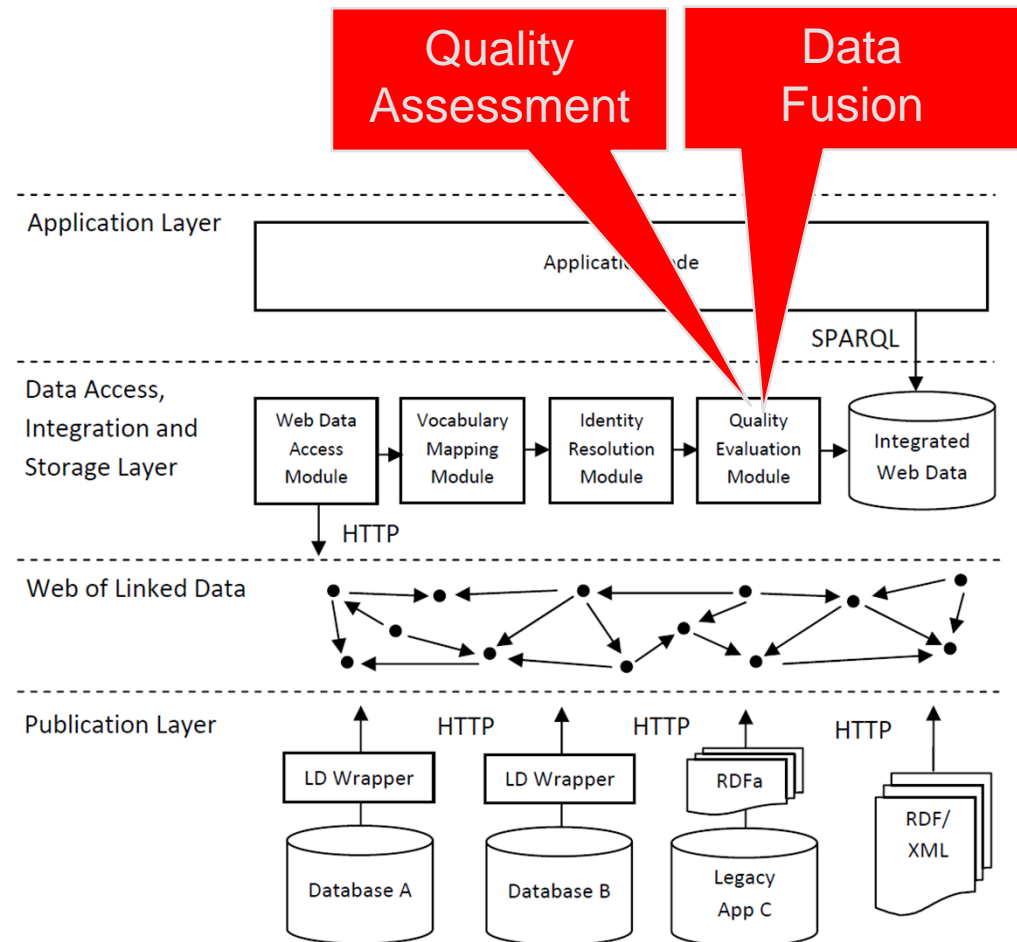
Silk Server

- Add missing links while consuming Linked Data
- Designed to work together with LDspider
- Talk about Silk Server at COLD workshop



WIQA Framework

- Allows you to filter Web data using different data quality assessment policies
- Will be extended towards Data Fusion



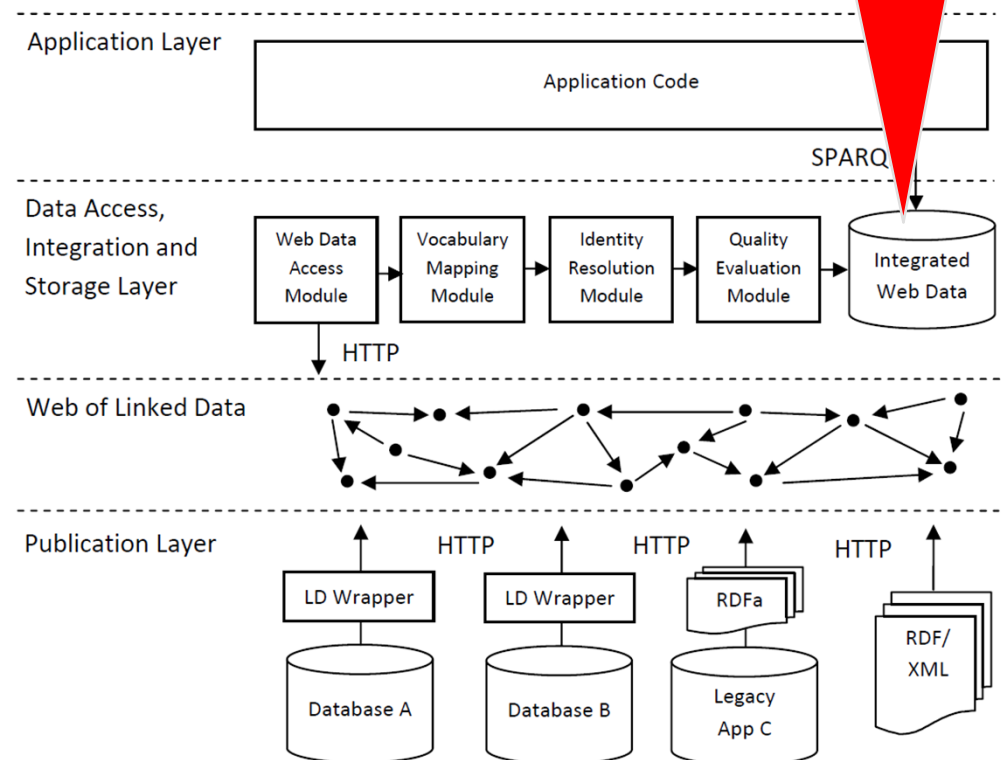
Named Graphs

- Extension of the RDF Data Model for representing meta-information about RDF Graphs
- Implemented by most SPARQL stores
- Used by many Linked Data applications for provenance tracking

- **Provenance vocabularies**

- Are compared by W3C Provenance XG
- Open Provenance Model is gaining traction

Representation of provenance meta-information



Conclusion

■ The Web of Linked Data is growing rapidly

- Active deployment communities exist in various domains
- Has exceeded the critical mass

■ Web search is evolving into query answering

- Search engines will increasingly rely on structured data from the Web

■ Next step: Linked Data within Enterprises

- alternative to data warehouses and EAI middleware
- advantages: schema-less data model, pay-as-you go data integration

■ You are looking for a topic for your PhD thesis?

- There are many exciting research challenges around consuming Linked Data
- Examples: Web-scale data integration, data quality assessment

Thanks!

References

- 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>
- 1st Workshop on Consuming Linked Data at ISWC 2010
<http://people.aifb.kit.edu/aha/2010/cold/>
- 3rd Linked Data on the Web Workshop at WWW 2010
<http://events.linkedata.org/ldow2010/>