Workshop On
Global Scientific Data Infrastructures: The Big Data Challenges
May 13th 2011, Capri, Italy

The Web of Linked Data

A global Data Infrastructure build on Web Architecture

Christian Bizer

Freie Universität Berlin

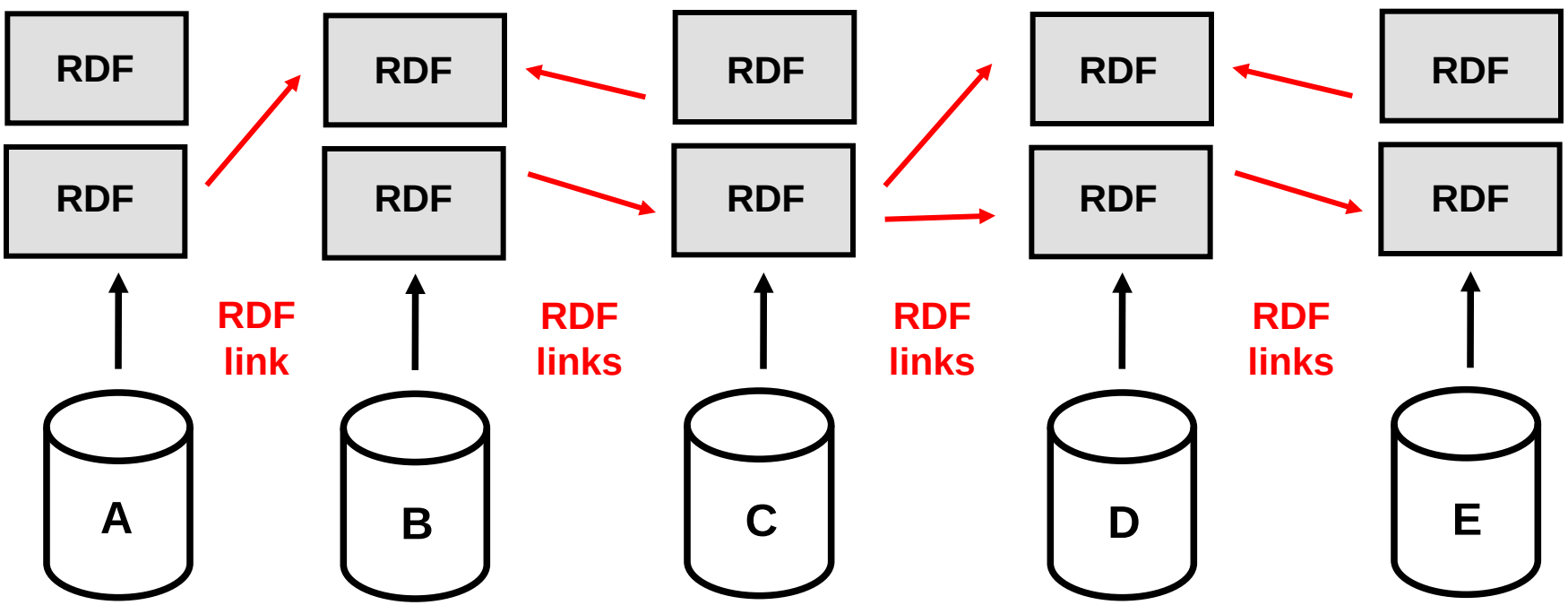
Germany

Outline

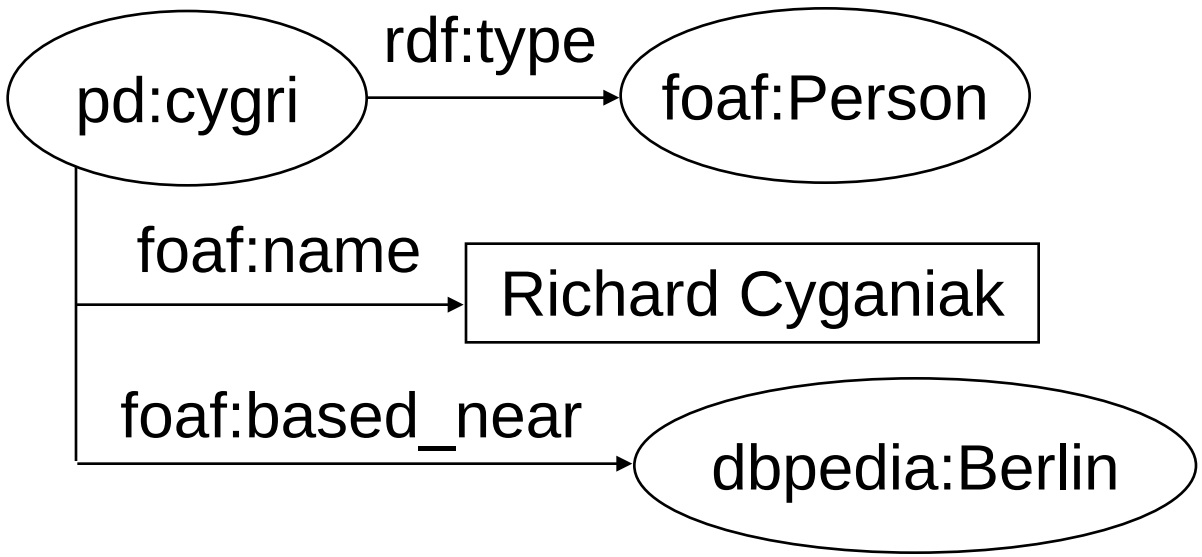
1. Linked Data
2. The Web of Linked Data
3. Splitting the Data Integration Effort

The Linked Data Principles

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

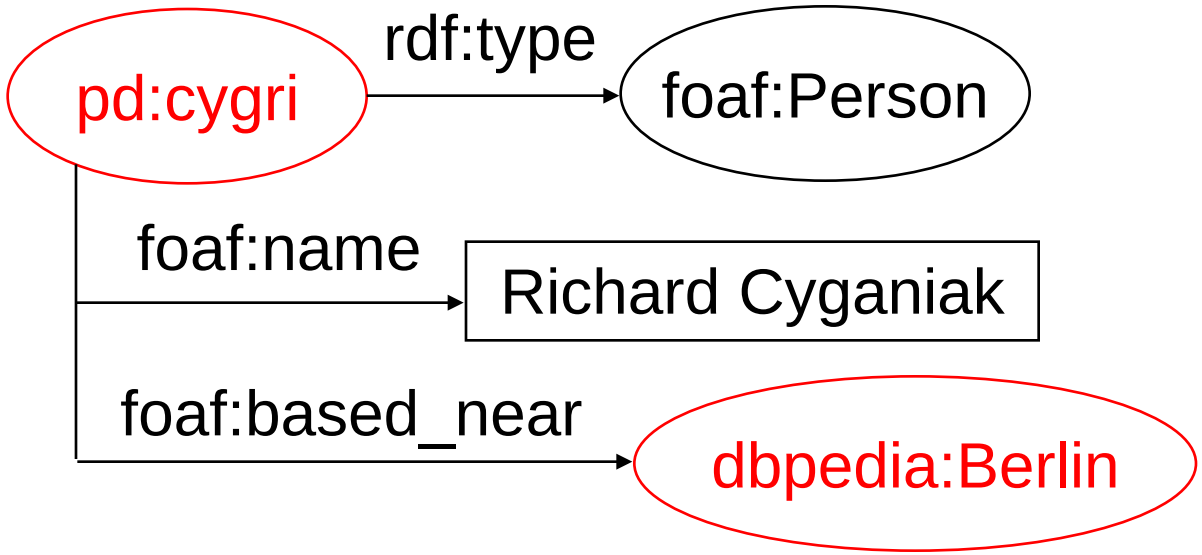


RDF Data Model



Flexible graph-based data model.

Data items are identified with HTTP URIs

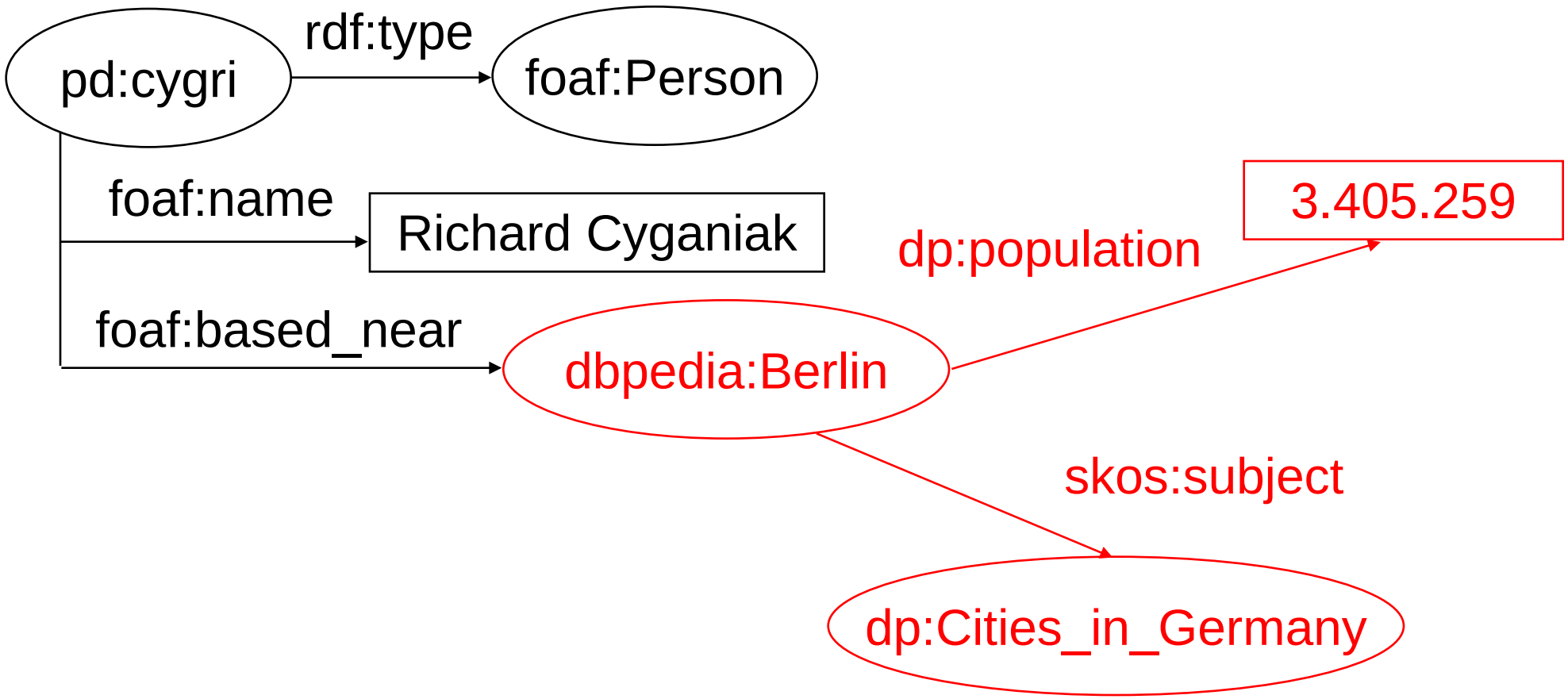


HTTP URIs take the role of global primary keys.

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

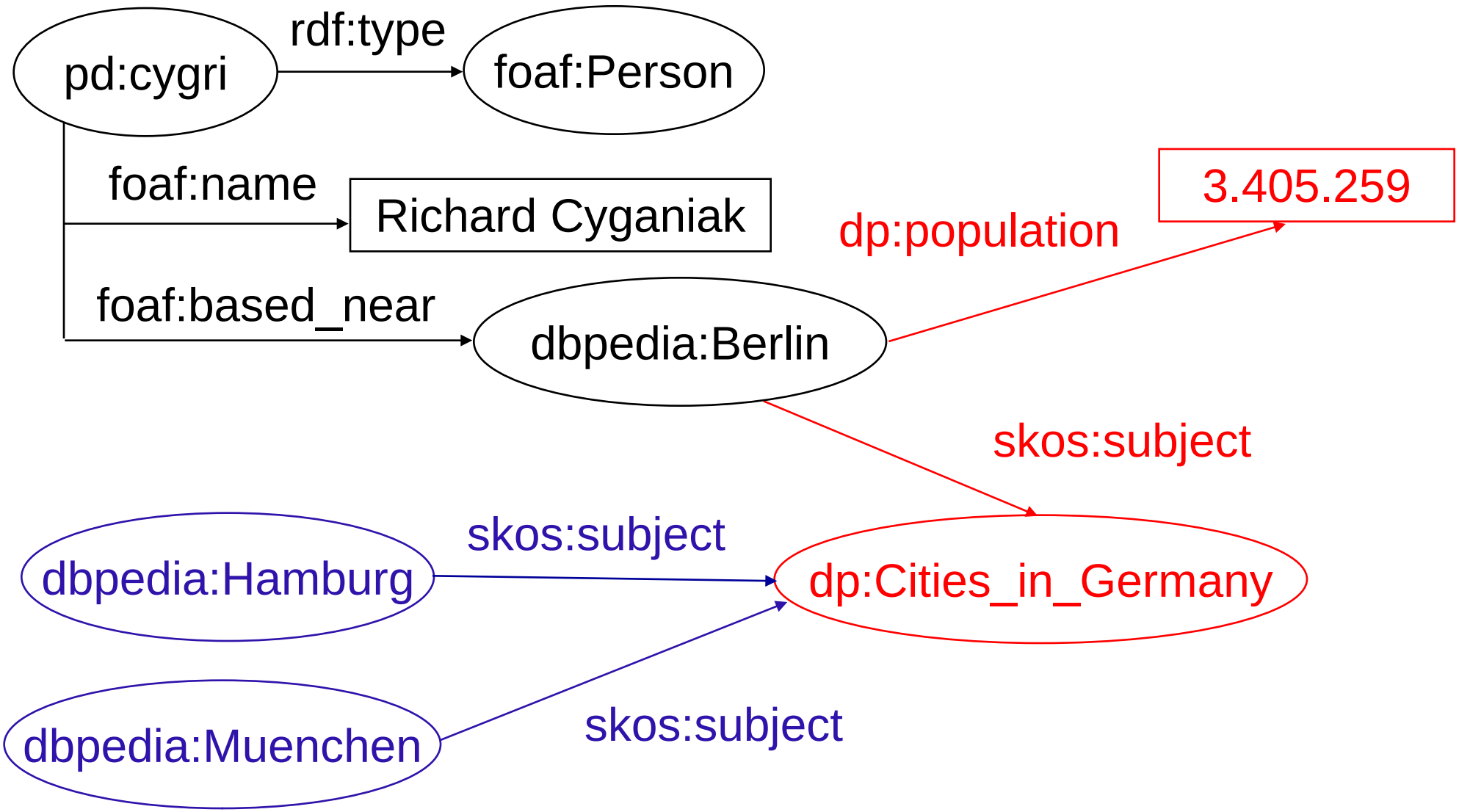
dbpedia:Berlin = http://dbpedia.org/resource/Berlin

Resolving URIs over the Web



The HTTP protocol brings together identification and retrieval.

Following Links deeper into the Web



Properties of the Web of Linked Data

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

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

Falcons

Object Concept Document

Chicago | Search Objects

Type
- **Any type**
- Abstraction
- Agent
- Athletic Activity
- Bull
- Cattle
- Concept
- Organisation
- Person
- Physical Entity
- Soccer Club
- Social Entity
- Spatial Thing
- Sports Team
- Subject
- Team

Objects **1** - **10** of **63,109** for your search **Chicago** (**1.25** seconds)

Chicago - Begriff
- label: **Chicago**
- type: Begriff

http://www4.wiwiss.fu-berlin.de/bookmashup/subject/**Chicago**

Chicago - City, Community
- label: **Chicago**
- comment: **Chicago** [;] (deutsch: Chikago) ist eine Stadt am Südwestufer des Michigansees im US-Bundesstaat Illinc USA. In der Agglomeration leben 9.443.356 Menschen (2005)"
- sameAs: http://www.rdfabout.com/rdf/usgov/geo/us/il/counties/cook_county/**chicago**
- image: 
- type: Community

http://dbpedia.org/resource/**Chicago**

chicago
- Title: **chicago**

http://www.deadjournal.com/interests.bml?int=**chicago**

Chicago Cubs players - Begriff
- label: **Chicago** Cubs players
- bevorzugter Name: **Chicago** Cubs players
- hat Oberbegriff: **Chicago** Cubs field personnel
- hat Oberbegriff: **Chicago** Cubs
- type: Begriff

http://dbpedia.org/resource/Category:**Chicago**_Cubs_players

People from **Chicago** - Begriff
- label: People from **Chicago**

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]

NG4J - Named Graphs API for Jena [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/pe

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/co

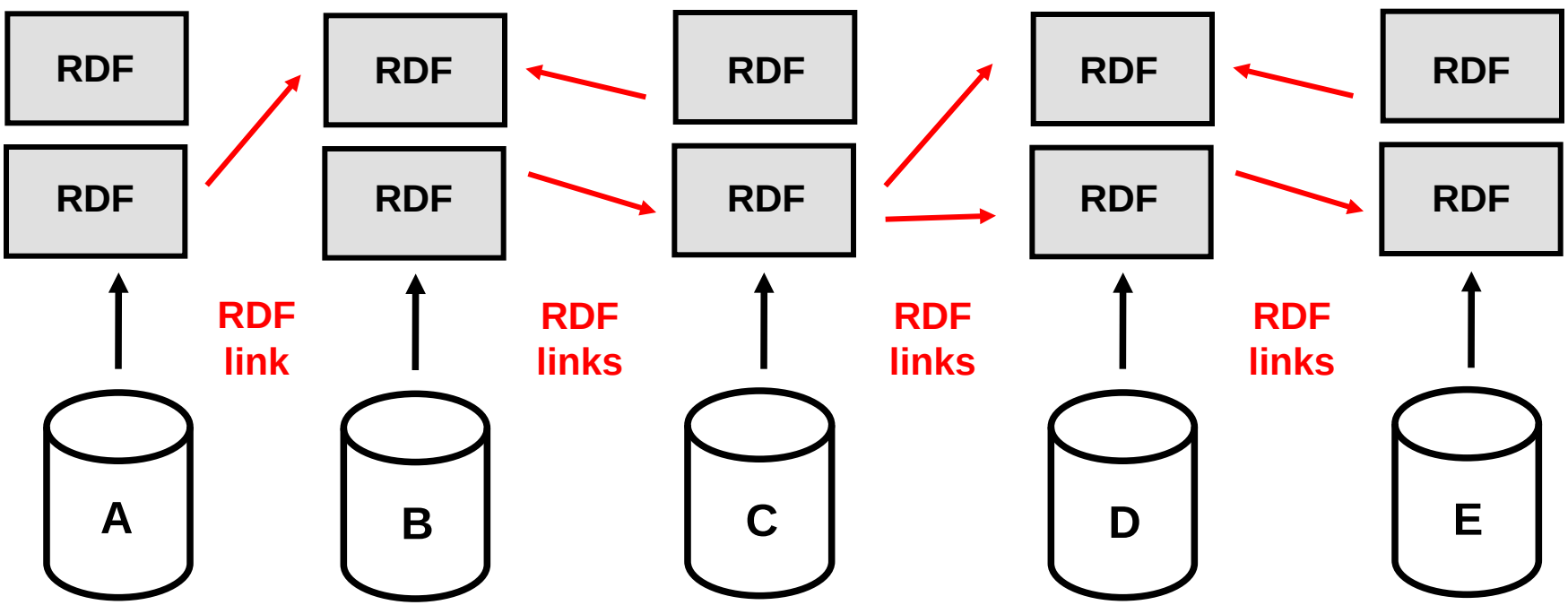
10 Chris Bizer 6 facts | 20
BOSS http://ebiquity.umbc

<- 1 2 ->

http://example.loc/doc

2. The Web of Linked Data

- Is this real?

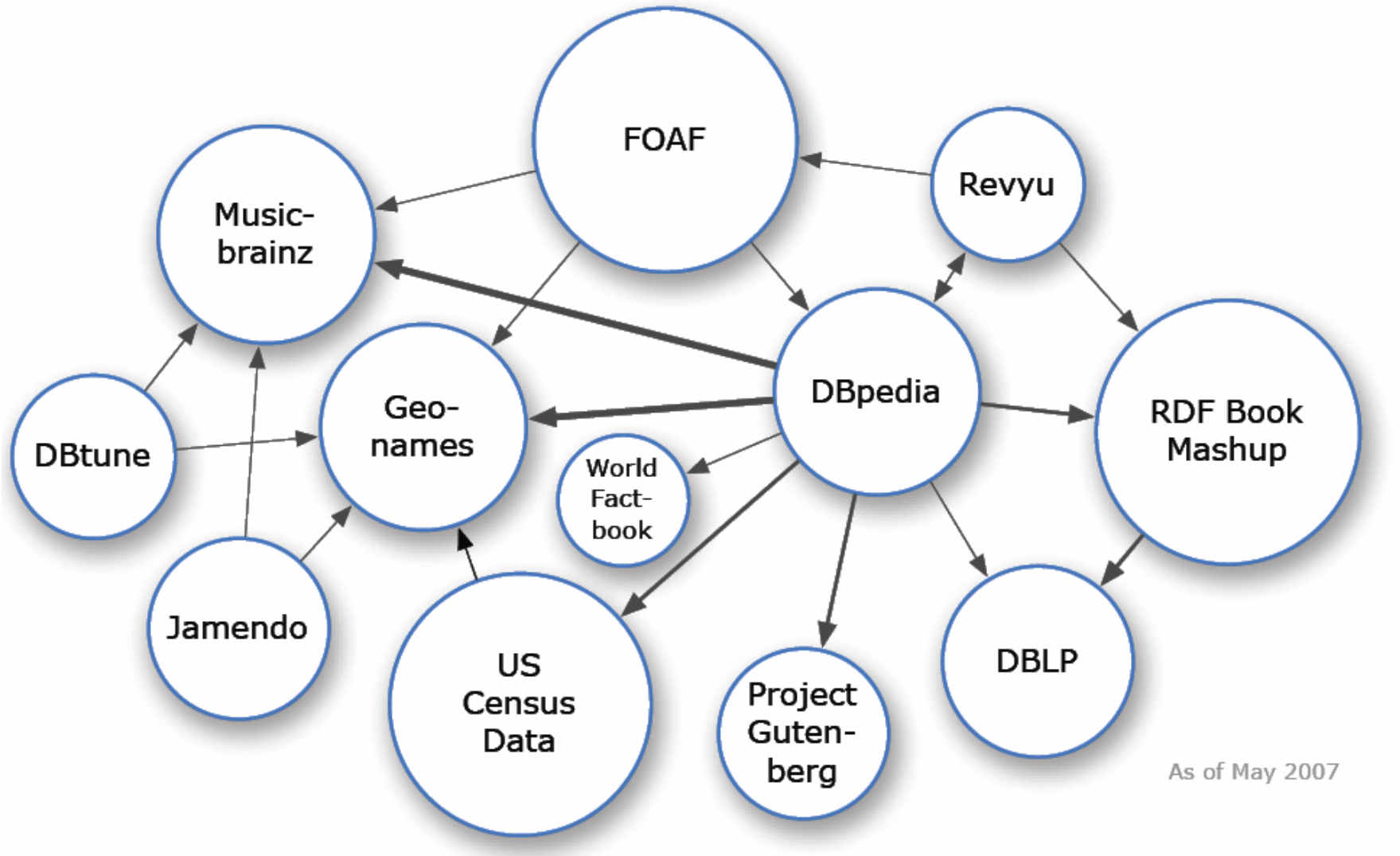


W3C Linking Open Data Project

LINKINGOPENDATA
W3C SWEO Community Project

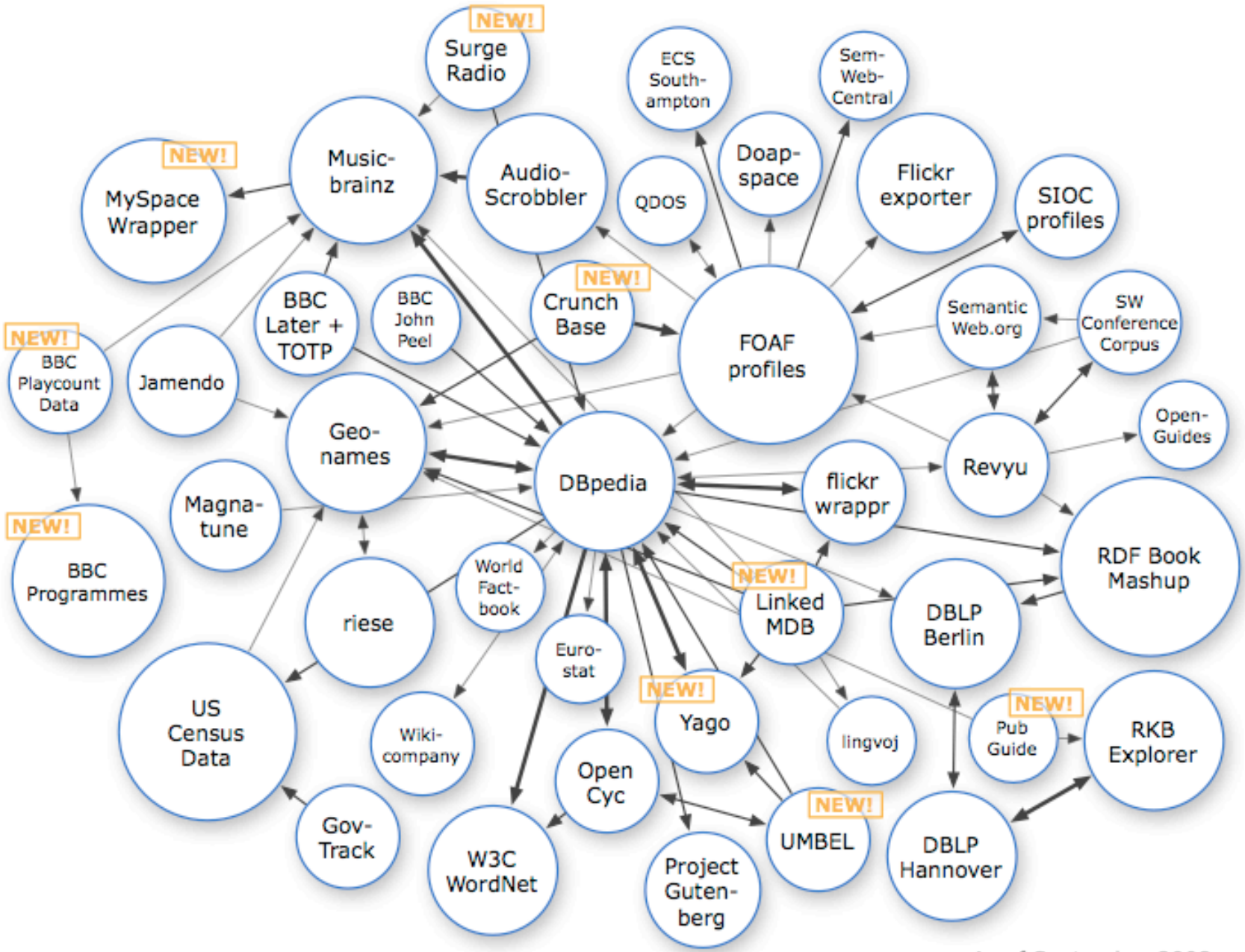
- **Grassroots community effort to**
  - publish existing open license datasets as Linked Data on the Web
  - interlink things between different data sources

LOD Datasets on the Web: May 2007

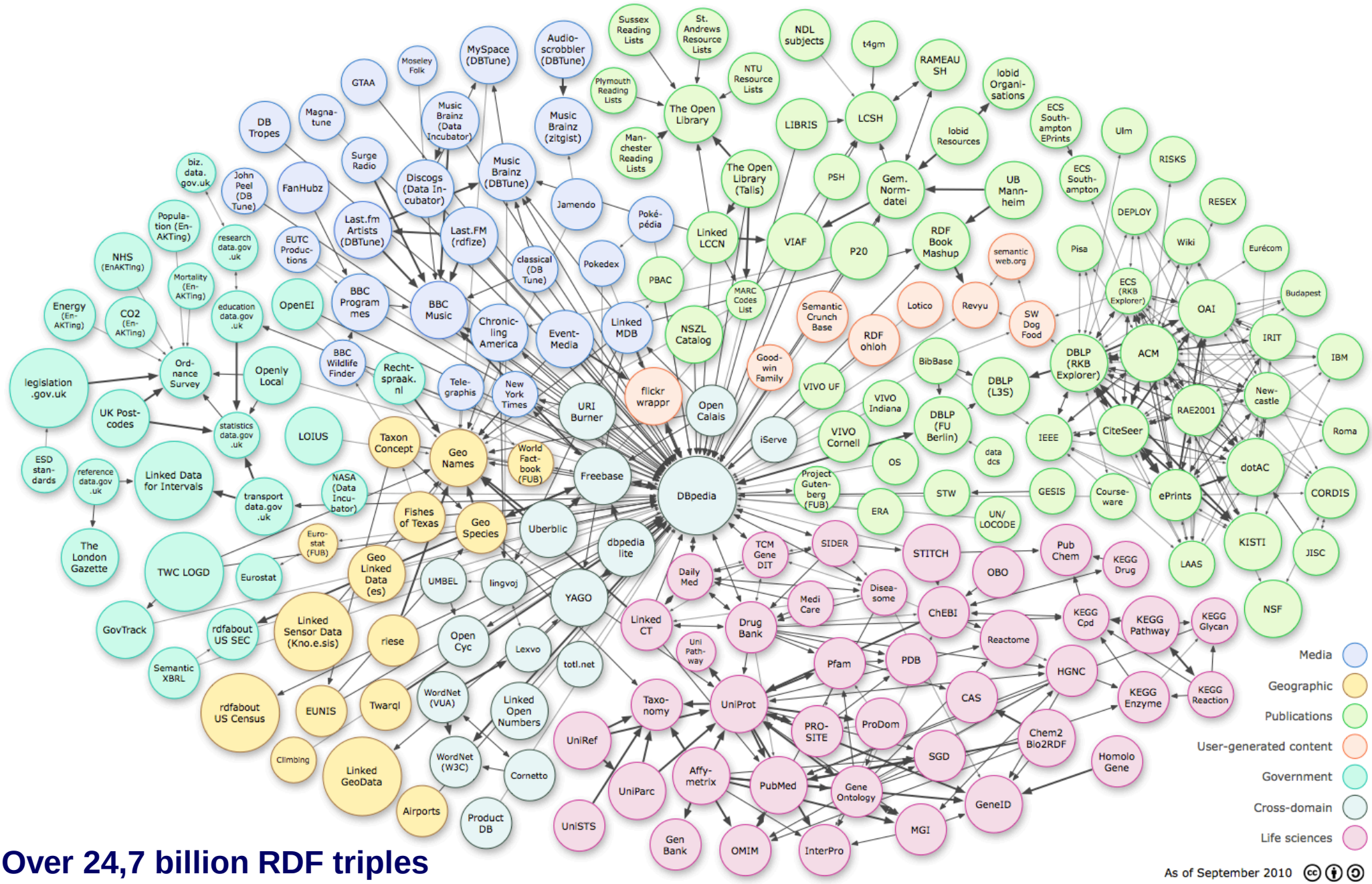


- **Over 500 million RDF triples**
- **Around 120,000 RDF links between data sources**

LOD Datasets on the Web: September 2008



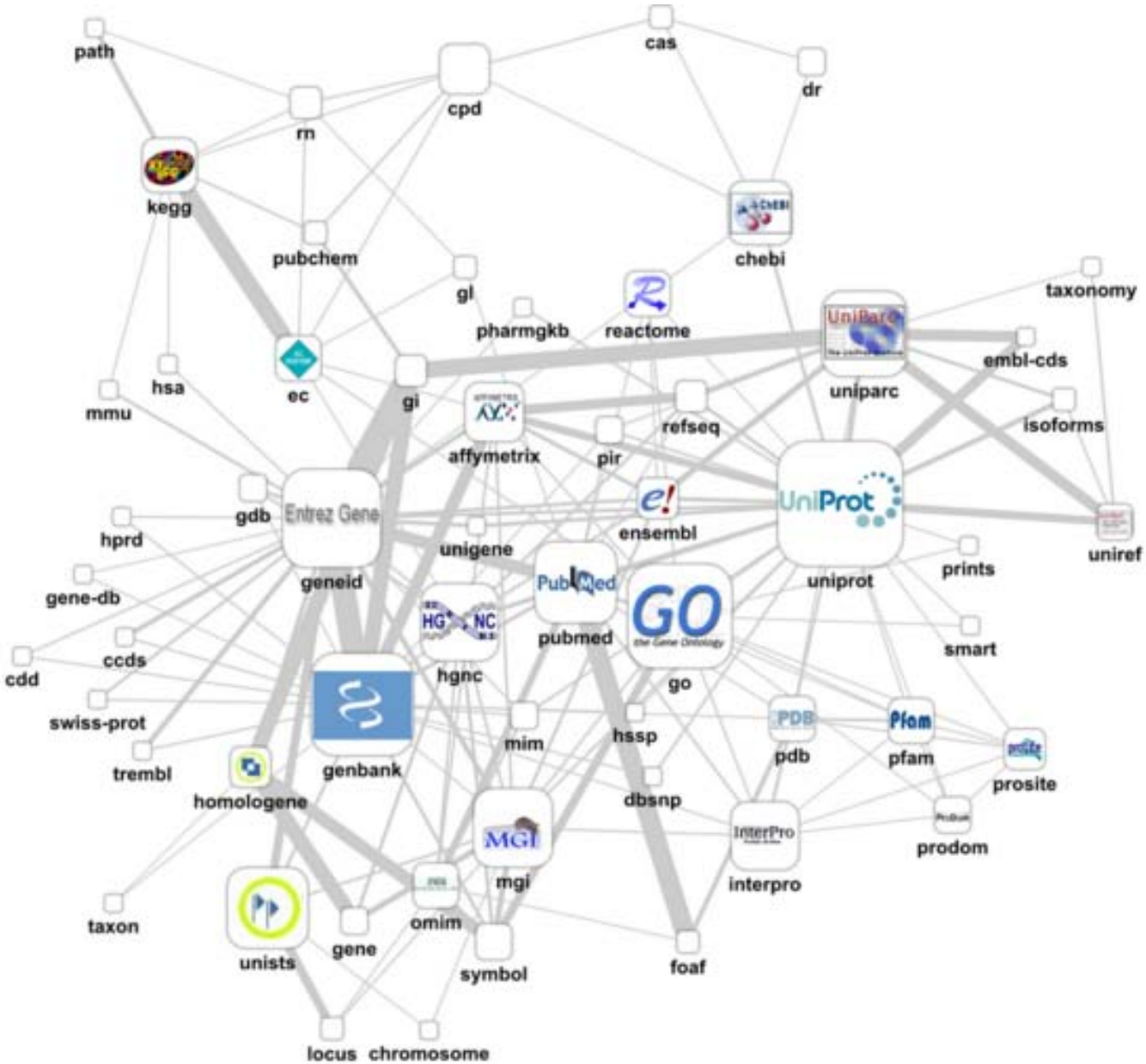
LOD Datasets on the Web: September 2010



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

Uptake in Life Sciences

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

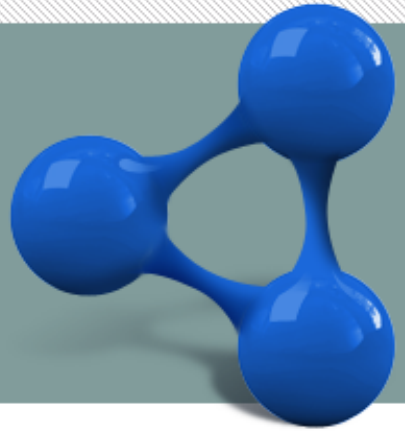


Uptake in the Libraries Community

- **Institutions publishing Linked Data**
  - Library of Congress (subject headings)
  - German National Library (PND dataset and subject headings)
  - Swedish National Library (Libris - catalog)
  - Hungarian National Library (OPAC and Digital Library)
  - The Europeana project just released data about 4 million artifacts
- **W3C Library Linked Data Incubator Group**

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

3. Splitting the Data Integration Effort

The Dataspace Vision

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

- **Properties of dataspaces**
  - require no upfront investment into a global schema
  - rely on pay-as-you-go data integration
  - give best effort answers to queries

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

Linked Data relies on the Pay-as-You-Go Idea

- for Identity Management
- for Vocabulary Management

Identity on the Web of Linked Data

Real world objects are identified with multiple URIs.

- **Everybody can say everything about everything.**
- **Cheap to set up.**

Wrapper around the DBLP bibliography

Linked Data website of our research group

http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer

http://www4.wiwiss.fu-berlin.de/is-group/resource/persons/Person4

Publish Identity Links on the Web

Identity Link

```
<http://www4.wiwiss.fu-berlin.de/is-group/resource/persons/Person4>
owl:sameAs
<http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer> .
```

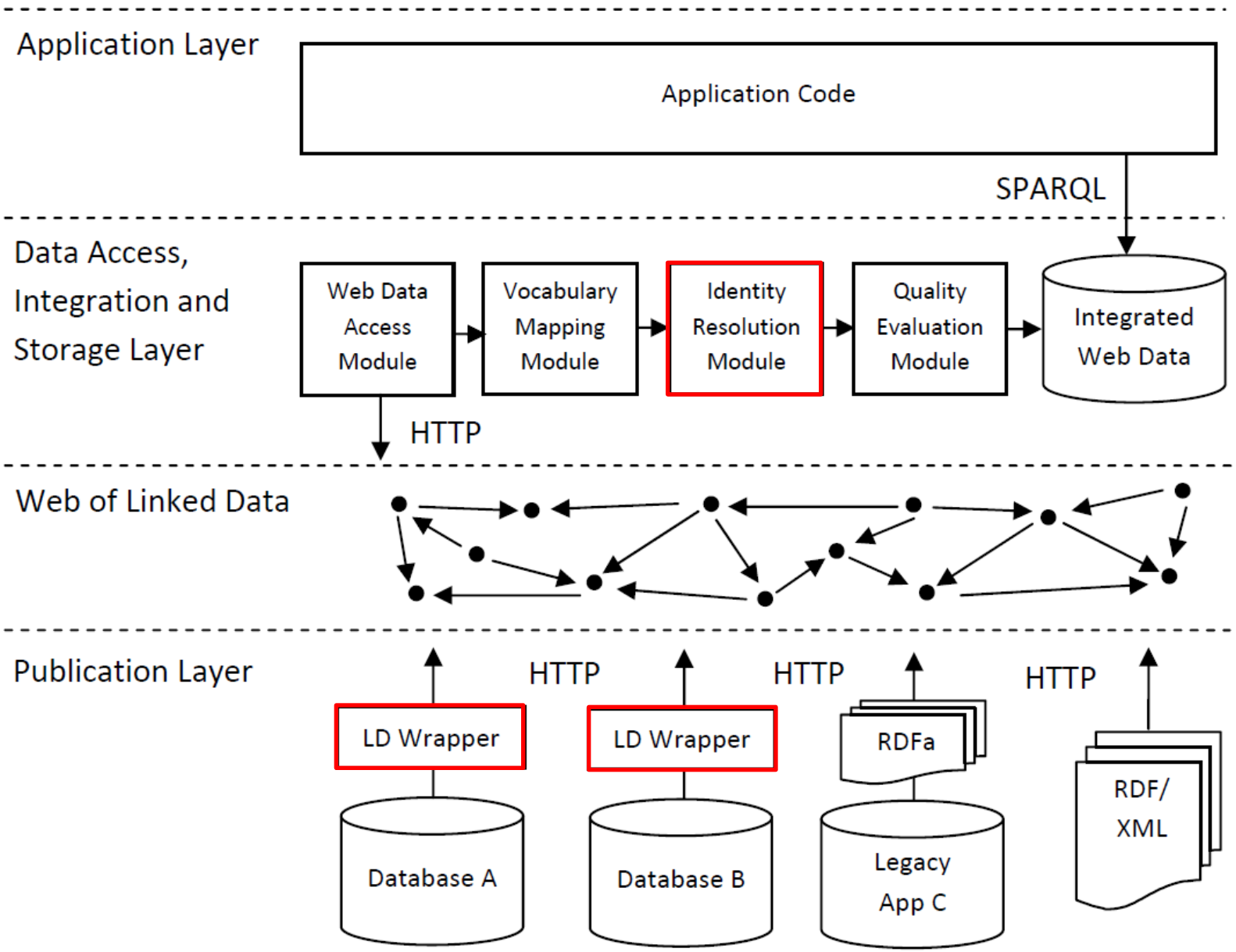
- **Pay-as-you-go Aspect**
  1. First: Just put a wrapper in front of your DB
  2. Later: You or somebody else invests effort into identity resolution
  3. Publishes the results as identity links on the Web

Effort Distribution between Publisher and Consumer

Consumer generates identity links

Effort Distribution

Publishers or third parties provides identity links



Vocabularies on the Web of Linked Data

Everyone can use whatever vocabulary or mixture of vocabularies that she likes to publish Linked Data on the Web.

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 Vocabulary Links on the Web

Vocabulary Link

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

- **Simple Mappings: RDFS, OWL**
  - rdfs:subClassOf, rdfs:subPropertyOf
  - owl:equivalentClass, owl:equivalentProperty
- **Complex Mappings: R2R**
  - provides value transformation functions
  - structural transformations
- **Pay-as-you-go Aspect**
  1. Use a mix of common vocabularies and proprietary terms
  2. You or somebody else publishes schema mappings afterwards

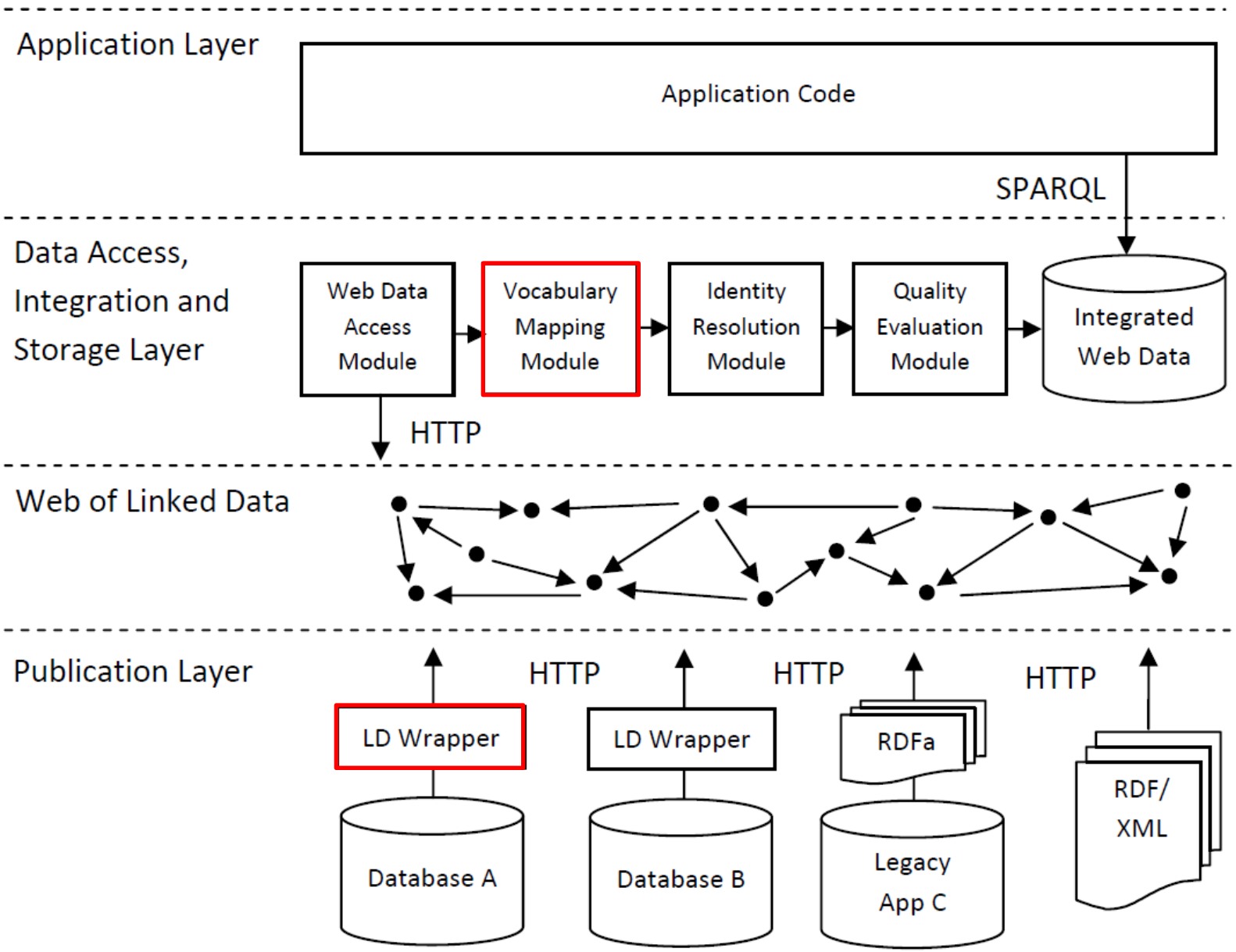
Effort Distribution between Publisher and Consumer

Consumer defines or data mines mappings

Effort Distribution

Publisher reuses vocabularies

Publisher or third party publishes mappings



Somebody-Pays-As-You-Go

The overall data integration effort is split between the data publisher, the data consumer and third parties.

- **Data Publisher**
  - publishes data as RDF
  - sets identity links
  - reuses terms or publishes mappings
- **Third Parties**
  - set identity links pointing at your data
  - publish mappings to the Web
- **Data Consumer**
  - has to do the rest
  - using record linkage and schema matching techniques

Conclusion

- **Linked Data realizes the dataspace vision on global scale and adds the social dimension to it.**
- **The Web of Linked Data is growing rapidly**
  - active deployment communities in different domains
  - has exceeded critical mass
- **Great playground for research and experimentation**
  - dataspace profiling
  - probabilistic and approximate schema mapping
  - data fusion, data quality, provenance
  - What will the user interfaces look like?
  - Will search engines turn into answer engines?

Thanks!

References

- Textbook: Tom Heath, Christian Bizer: *Linked Data: Evolving the Web into a Global Data Space.* http://linkeddatabook.com/
- 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
- 4th Linked Data on the Web Workshop at WWW 2011 http://events.linkeddata.org/ldow2011/
- 1st Workshop on Consuming Linked Data at ISWC 2010 http://people.aifb.kit.edu/aha/2010/cold/