International Conference on Database Theory (ICDT 2014)
Athens, Greece, 25.3.2014

Invited Lecture

Search Joins with the Web

Prof. Dr. Christian Bizer

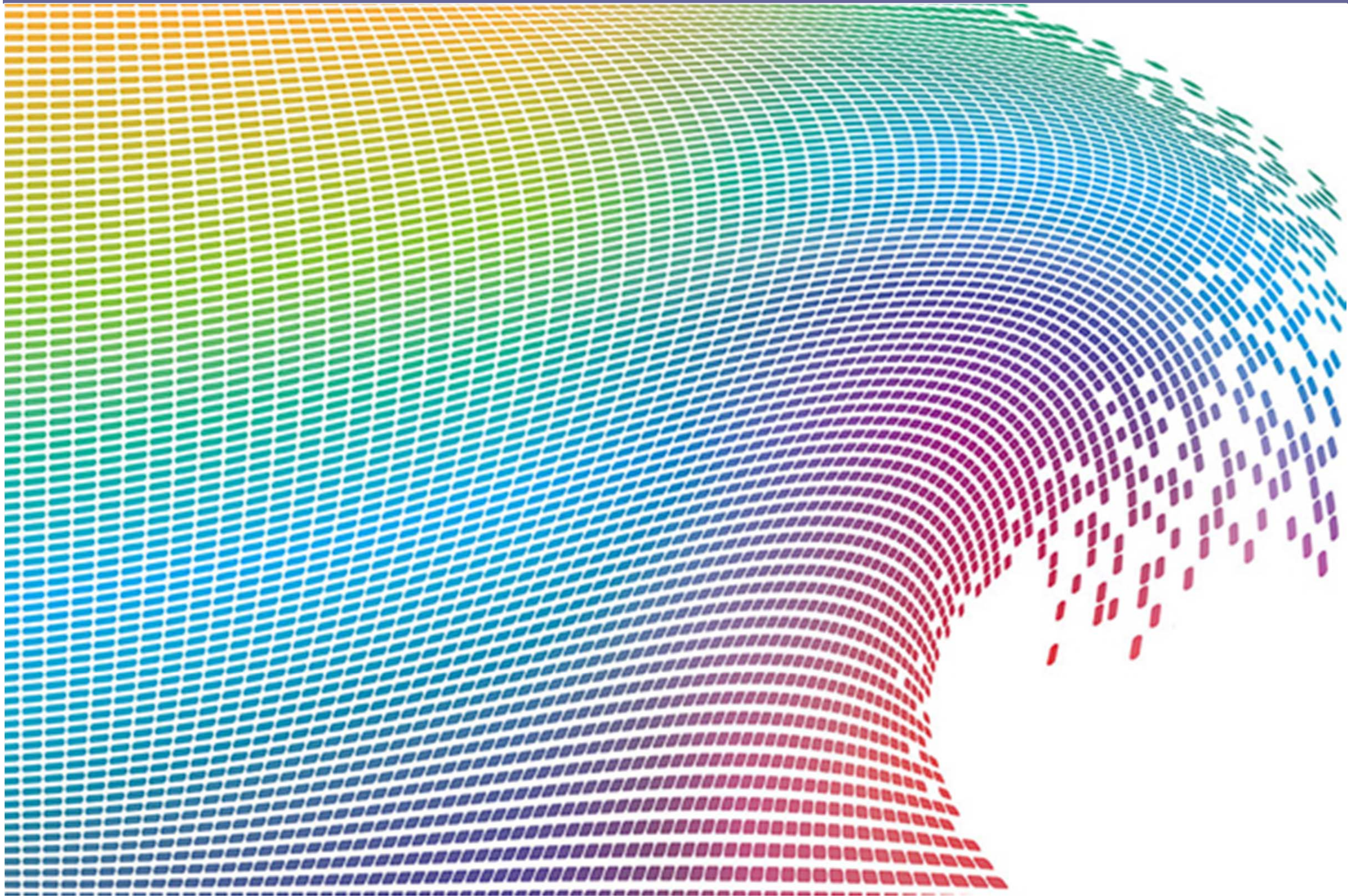
UNIVERSITÄT MANNHEIM

Outline

A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web.

1. Motivation and Definition
2. Profile of the available Web Data
3. Feasibility of Search Joins
4. Table Relevance

Deluge of Structured Data on the Web



Relational HTML Tables

In corpus of 14B raw tables, 154M are “good” relations (1.1%).

Germany Phone Code: 49
Berlin Area Code: 30
Berlin Dial Code: +49 30

City	Area Code	Dialing Code
Aachen	241	+49 241
Augsburg	821	+49 821
Bergisch Gladbach	2202	+49 2202
Berlin	30	+49 30
Bielefeld	521	
Bonn	228	
Bottrop	2041	

Largest German cities

German cities ranked 1 to 100 | German cities ranked 101 to 200 | German cities ranked 201 to 300 |

Rank	City	State	Population
1	Berlin	Berlin	3,275,000
2	Hamburg	Hamburg	1,686,100
3	München	Bavaria	1,185,400
4	Köln	Northrhine-Westfalia	965,300
5	Frankfurt	Hessen	648,000
		Northrhine-Westfalia	588,800
		Northrhine-Westfalia	587,600
		Baden-Württemberg	581,100
		Northrhine-Westfalia	568,900
		Bremen	527,900

Germany - Largest Cities

	Name	Population	Latitude/Longitude
1	Berlin, Berlin	3,426,354	52.524 / 13.411
2	Hamburg, Hamburg	1,739,117	53.575 / 10.015
3	Munich, Bavaria	1,260,391	48.137 / 11.575
4	Cologne, North Rhine-Westphalia	963,395	50.933 / 6.95
5	Frankfurt am Main, Hesse	650,000	50.116 / 8.684
6	Essen, North Rhine-Westphalia	593,085	51.457 / 7.012
7	Stuttgart, Baden-Württemberg	589,793	48.782 / 9.177
8	Dortmund, North Rhine-Westphalia	588,462	51.515 / 7.466
9	Duesseldorf, North Rhine-Westphalia	573,057	51.222 / 6.776
10	Bremen, Bremen	546,501	53.075 / 8.808

- Cafarella, et al.: WebTables: Exploring the Power of Tables on the Web. VLDB 2008.

HTML-embedded Data on the Web

Several million websites semantically markup the content of their HTML pages.

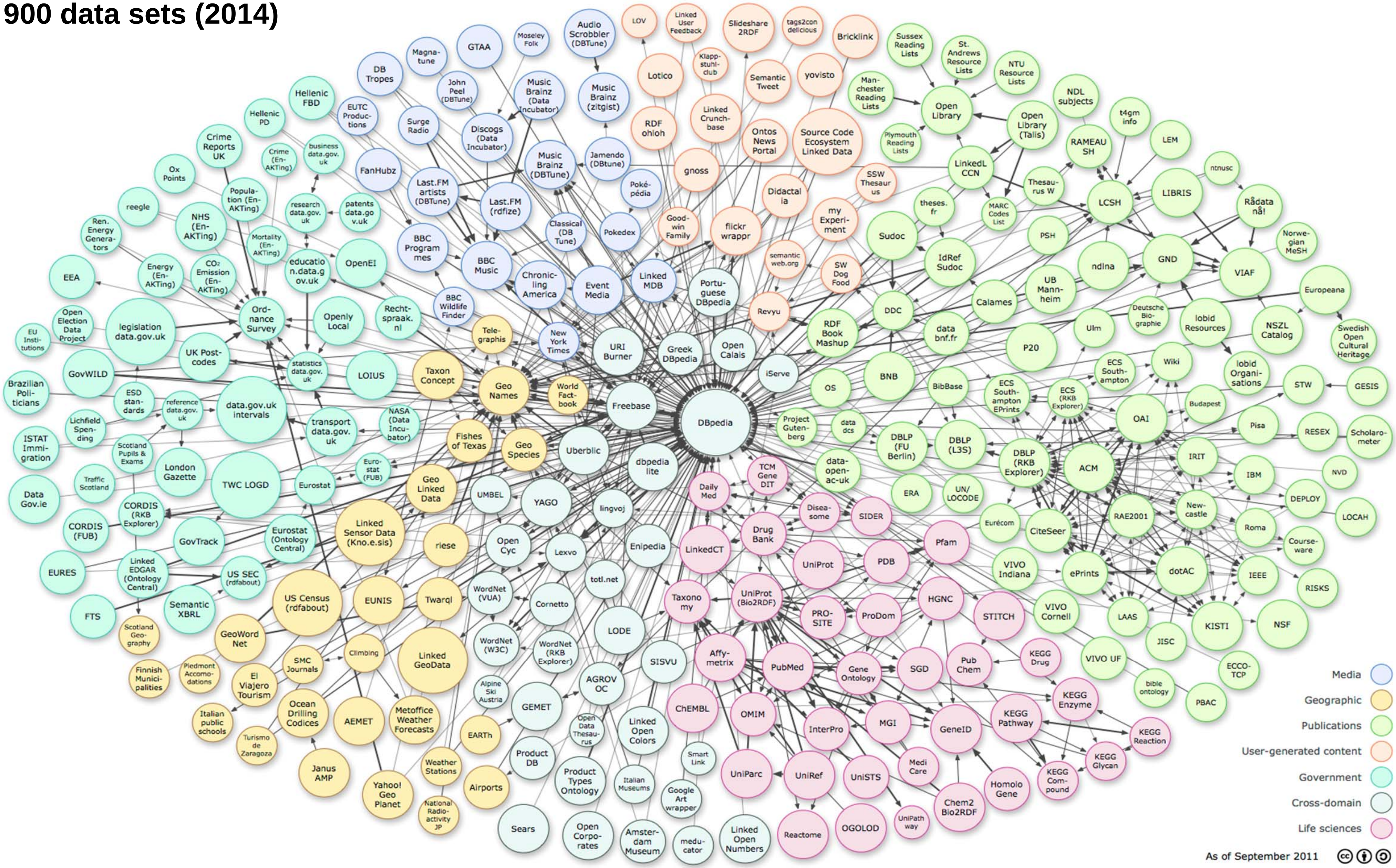
Markup Syntaxes

- Microformats
- RDFa
- Microdata

Data Consumers

Linked Data on the Web

~ 900 data sets (2014)



Data Portals

Several 100.000 datasets are available via data portals.

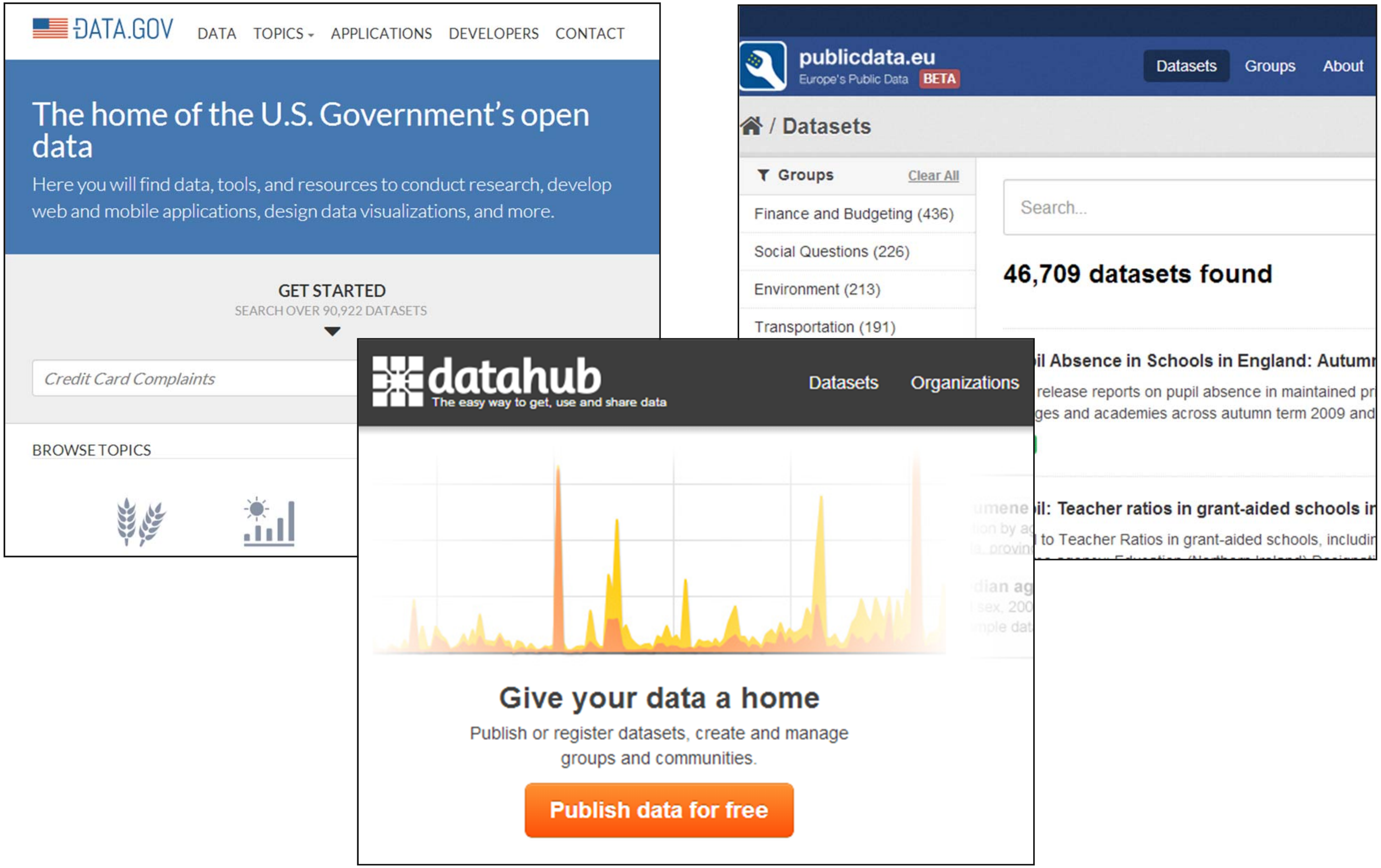
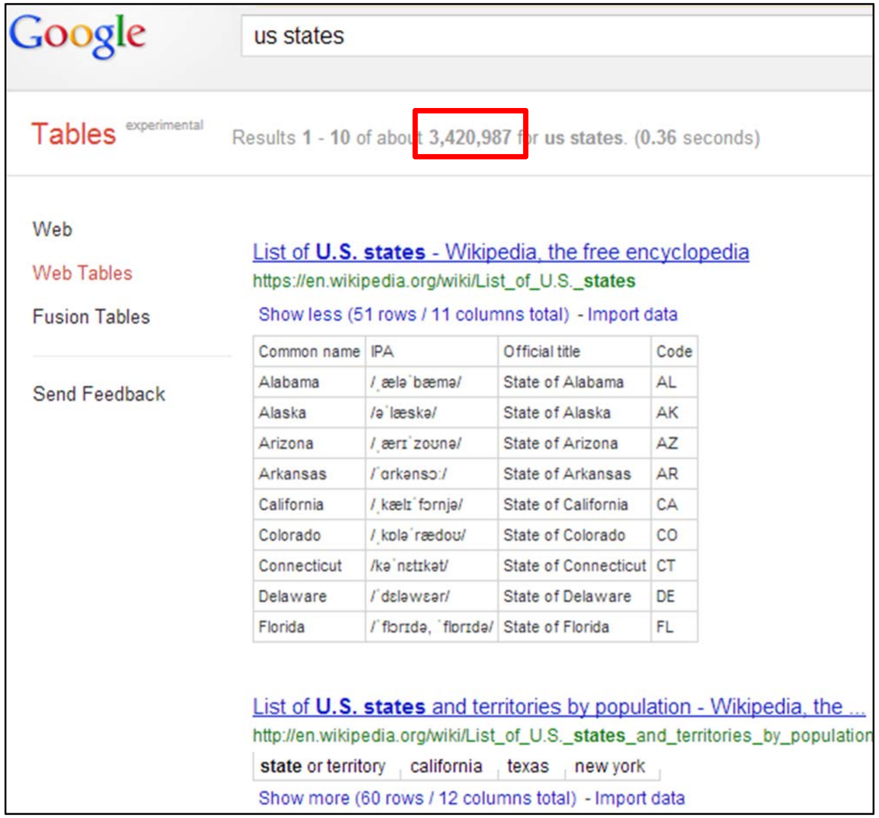


Table Search

Given some keywords describing the user's information need, generate ranked list of relevant tables.

- Example: Google Table Search
  – http://research.google.com/tables
- Problem:
  The user is left alone
  with the integration.
  – using some tool
  – doing cut@paste



- Venetis, et al.: Table Search Using Recovered Semantics. VLDB, 2010.
- Pimplikar, Sarawagi: Answering table queries on the web using column keywords. VLDB 2012.

Goal 1: Extend Local Table with Single Attribute

Given a local table, a search attribute, and keywords describing the extension attribute, add the extension attribute to the table and fill it with data from the Web.

Region

„GDP per Capita“

No.	Region	Unemployment
1	Alsace	11 %
2	Lorraine	12 %
3	Guadeloupe	28 %
4	Centre	10 %
5	Martinique	25 %
…	…	…

+

GDP per Capita
45.914 €
51.233 €
19.810 €
59.502 €
21,527 €
…

Goal 2: Extend Local Table with Many Attributes

Given a local table, a search attribute, add all attributes to the local table that can be filled beyond a density threshold.

Region

density >= 0.8

No.	Region	Unemp. Rate
1	Alsace	11 %
2	Lorraine	12 %
3	Guadeloupe	28 %
4	Centre	10 %
5	Martinique	25 %
…	…	…

+

GDP per Capita	Population Growth	Overseas departments	…
45.914 €	0,16 %	No	…
51.233 €	-0,05 %	No	…
19.810 €	1,34 %	Yes	…
59.502 €	NULL	NULL	…
NULL	2,64 %	Yes	…
…	…		…

Prototype: RapidMiner Linked Open Data Extension

Features

1. Extend local table with additional attributes from a Linked Data source
2. Add more attributes by following RDF links into other Linked Data sources
3. Mine extended tables using all RapidMiner features

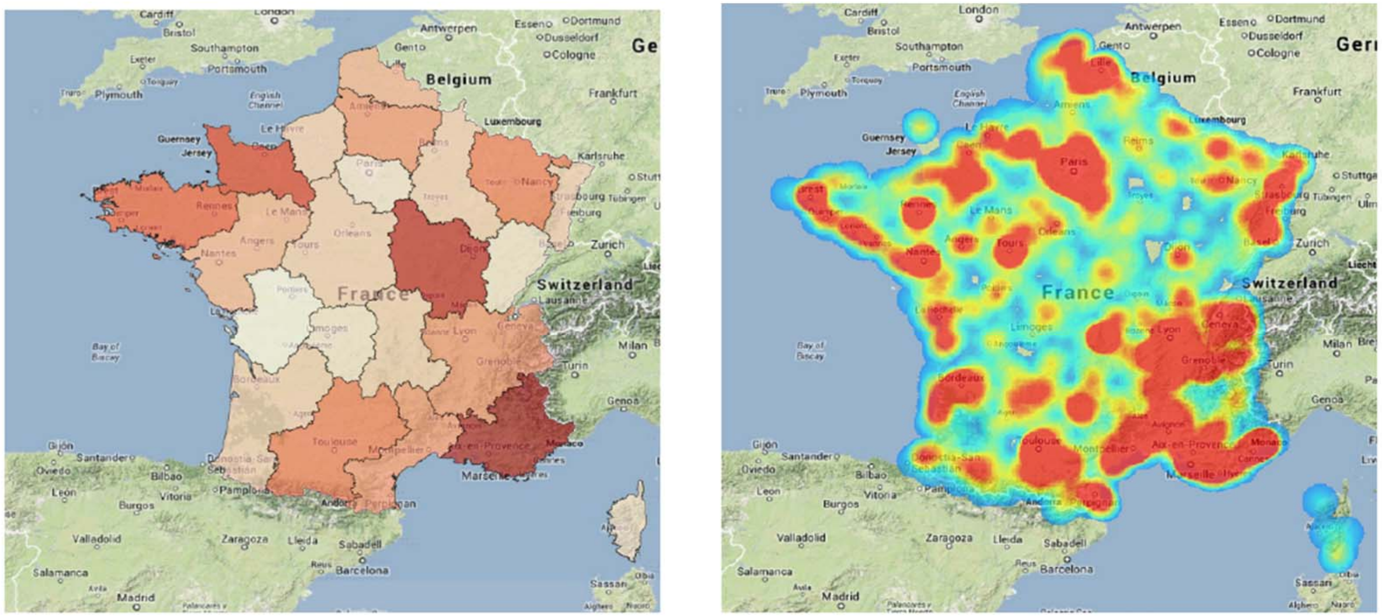
Input Table **Link** **Additional Attributes**

Row No.	region	unempl...	region_uri	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...	name_uri_...
18	Bretagne	0.097	http://wifo5-0	71136.200	1097.057	1.220	0.790	15536.500	13490	924	888.400	3704	248
19	Pays de la L	0.098	http://wifo5-0	82131.800	756.007	0.980	0.840	15720.600	15631	3494.400	733	4842	338
16	Limousin	0.100	http://wifo5-0	15703.300	124.380	0.780	0.280	16510.400	3282	1248.700	1017.700	1528	78
15	Rhône-Alpes	0.102	http://wifo5-0	158329	3896.363	1.880	0.890	17003	56881	28175.900	807.700	9428	455
14	Auvergne	0.102	http://wifo5-0	29729.200	689.348	1.580	0.260	16396.800	6450	1009.400	951.900	2374	151
11	Centre	0.107	http://wifo5-0	59502	868.893	1.140	0.340	16880.700	12992	13745.600	761.100	3838	317
12	Basse-Norm	0.108	http://wifo5-0	32224.600	297.951	0.910	0.280	15678.300	7117	3085.400	760.100	2148	146
13	Bourgogne	0.108	http://wifo5-0	37778.200	358.552	0.890	0.150	16654.500	8742	338.700	848.900	3206	178
21	Midi-Pyrénée	0.110	http://wifo5-0	64525	2282.610	2.180	1.160	15728.200	12598	7964	784.500	4820	304
3	Franche-Coi	0.110	http://wifo5-0	26343.400	529.876	1.390	0.390	16445.900	6989	689.800	758.100	1649	134
22	Île-de-France	0.110	http://wifo5-0	469046.700	14364.372	3.140	0.620	20912	51984	5914	660.200	24309	448
1	Alsace	0.112	http://wifo5-0	45914	692.110	1.350	0.700	17112.900	11960	3563.700	883.100	2597	119
23	Aquitaine	0.112	http://wifo5-0	74099.700	1146.621	1.070	0.930	16118.900	15200	5894	802.800	6088	349
17	Poitou-Chari	0.113	http://wifo5-0	37880.700	305.333	0.760	0.620	15766.100	7850	138.500	698.700	2490	239
25	Corse	0.114	http://wifo5-0	5686.200	13.209	0.900	0.970	14775.800	858	448	921.100	1115	44
2	Lorraine	0.122	http://wifo5-0	51233.500	546.504	0.920	0.160	16035.400	17315	9565.700	824.400	3640	192
10	Haute-Norm	0.124	http://wifo5-0	43191.300	600.947	0.950	0.240	16464.700	14035	11481.400	650	2418	155
6	Champagne	0.125	http://wifo5-0	33075	237.687	0.560	-0.100	15911.500	8134	5393.100	740.200	2065	124
24	Provence-Alı	0.131	http://wifo5-0	117460.300	2099.396	1.650	0.870	16582.400	27335	5690	813.200	14913	469

http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/

Resulting Correlations

- Overseas Department (positive)
- Population growth (positive)
- Fast food restaurants (positive)
- Police stations (positive)
- Hospital beds/inhabitants (negative)
- GDP (negative)
- Energy consumption (negative)



(a) Unemployment by region (b) Heat map of police stations

Linked Data Sources used: Eurostat and DBpedia

Definition: Search Join

A Search Join is a join operation which extends a local table with additional attributes based on the large corpus of structured data that is published on the Web.

- Input:
  1. Corpus of heterogeneous Web tables
  2. Query table
  3. Search attribute definition
  4. Extension attribute(s) definition
     - Single attribute case: keyword describing extension attribute
     - Multiple attributes case: density threshold
- Output:
  – Query table augmented with additional attribute(s)

Search Joins in SQL

```
SELECT city.*, web.population
FROM city SEARCH JOIN web ON city.name;
```

```
SELECT city.name, web.*(0.9)
FROM city SEARCH JOIN web ON city.name;
```

Elements of a Search Join

$$R = c\left(m\left(q, s_{q,s,a}(T_{web})\right)\right)$$

s() : Search operator determines the set of the top-k relevant Web tables. (Web tables which are beneficial join partners)

m() : MultiJoin operator performs a series of left-outer joins between the query table and all tables in the input set.

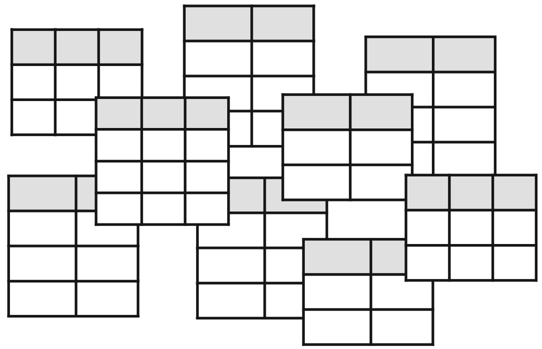
c() : Consolidation operator merges corresponding attributes and fuses attribute values in order to return a concise result table containing high-quality data.

The Search Operator

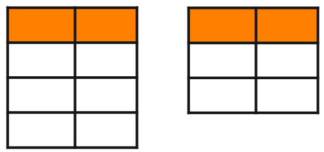
The Search operator determines the set of relevant Web tables.

$$T_r = s_{q,s,a}(T_{web})$$

- Input
  - T_{Web} = set of Web tables
  - q = query table
  - s = search attribute
  - a = attribute description



- Output
  - T_r = set of relevant Web Tables



Multi-Join Operator

The MultiJoin operator performs a series of left-outer joins between the query table and all tables in the input set.

- Input

  q = query table

  T_r = set of relevant Web tables

$$t_e = m(q, T_r)$$

- Output

  t_e = extended query table

No.	Region	Unemploy	Unemploy	GDP	GDP per C
1	Alsace	11 %	NULL	45.914 €	45.000 €
2	Lorraine	12 %	NULL	51.233 €	NULL
3	Guadeloupe	28 %	NULL	NULL	19.000 €
4	Centre	10 %	9.4 %	NULL	59.500 €

Consolidation Operator

The consolidation operator merges corresponding attributes and fuses attribute values in order to return a concise result table containing high-quality data.

$$t_r = c(t_e)$$

- Input

  t_e = extended query table

- Output

  t_r = result table

No	Region	Unemploy	GDP
1	Alsace	11 %	45.914 €
2	Lorraine	12 %	51.233 €
3	Guadelo upe	28 %	19.000 €
4	Centre	10 %	59.500 €

- Might employ various sources of attribute correspondences.
- Might employ various conflict resolution functions.

Data Model for Representing Web Data

- Entity-Attributes-Tables
  - One entity per row
- Subject Attribute
  - name of the entity
  - string, no number or other data type
  - relatively unique values

Rank	Film	Studio	Director	Length
1.	Star Wars –Episode 1	Lucasfilm	George Lucas	121 min
2.	Alien	Brandwine	Ridley Scott	117 min
3.	Black Moon	NEF	Louis Malle	100 min

Data Model Details

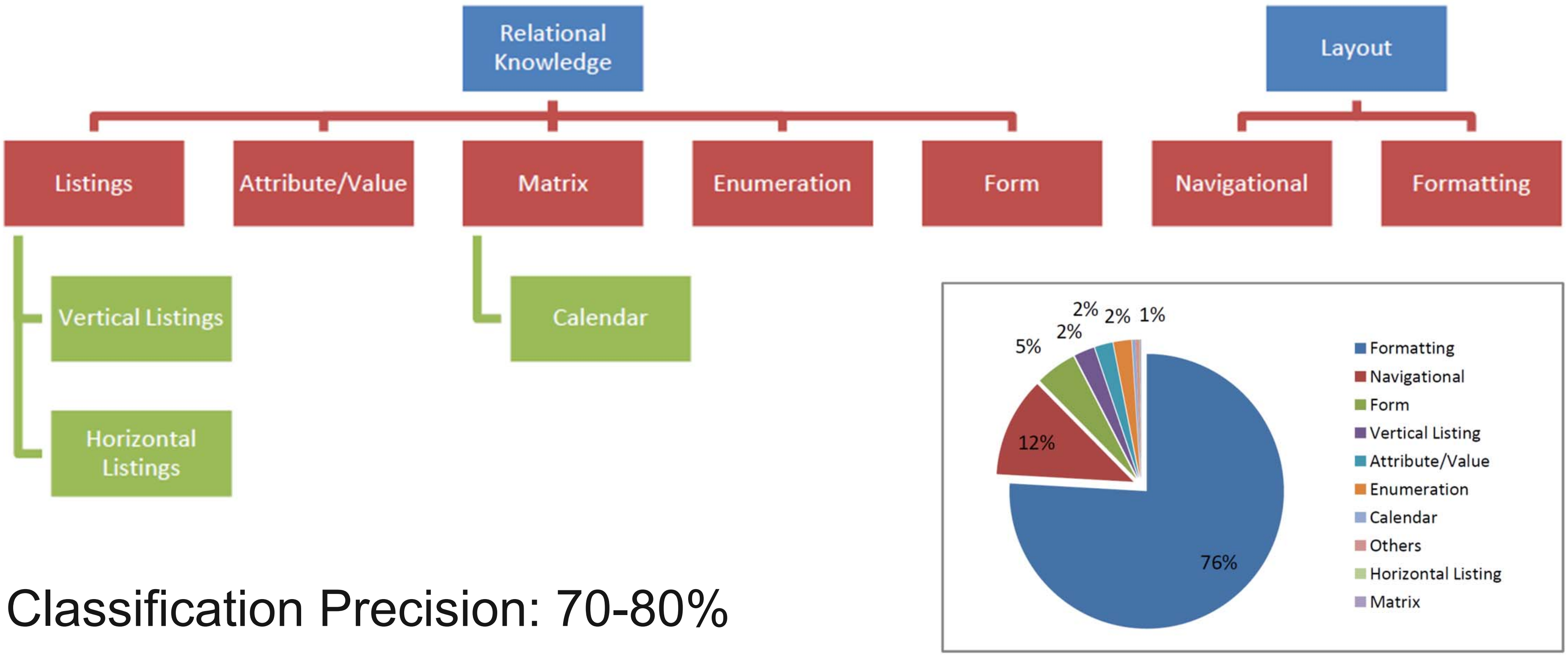
- Table Meta-Information
  - Provenance: Source URL
  - Table Context: Text around the table
- Attribute Headers
  - Attribute name
  - Unit of measurement
  - Date
- Data types
  - String, Number, Date, Geo Coordinates
  - Lists
  - URI Reference

2. Profile of the available Web Data

HTML Tables

In corpus of 14B raw tables, 154M are “good” relations (1.1%).

Cafarella (2008)



Classification Precision: 70-80%

- Cafarella, et al.: WebTables: Exploring the Power of Tables on the Web. VLDB 2008.
- Crestan, Pantel: Web-Scale Table Census and Classification. WSDM 2011.

Subject Attribute and Header Detection

- **Subject Attribute Detection** (Ventis)
  - Simple heuristic approach (Accuracy: 83%)
    - scan columns from left to right
    - take first column that is not a number or a date
  - SVM Classifier (Accuracy: 94%)
    - fraction of cells with unique content
    - variance in the number of tokens in each cell
    - column index from the left
    - ….
- **Header Detection** (Pimplikar)
  - one header row: 60%
  - two or more header rows: 22%
  - no header: 18%

- Ventis, et al.: Recovering Semantics of Tables on the Web. VLDB 2011.
- Pimplikar, Sarawagi: Answering table queries on the web using column keywords, VLDB 2012.

The Common Crawl

Web Data Commons – Web Tables Corpus

- Large corpus of relational Web tables for public download
- extracted from Common Crawl 2012 (3.3 billion pages)
- 147 million relational tables
  - selected out of 11.2 B raw tables (1.3%)
  - download includes the HTML pages of the tables (1TB zipped)

- Table Statistics

	Min	Max	Avg	Median
Attributes	2	2,368	3.49	**3**
Data Rows	1	70,068	12.41	**6**

- Heterogeneity: Very high.

- http://webdatacommons.org/webtables/

Web Data Commons – Web Tables Corpus

- Attribute Statistics

Attribute	#Tables
name	4,600,000
price	3,700,000
date	2,700,000
artist	2,100,000
location	1,200,000
year	1,000,000
manufacturer	375,000
counrty	340,000
isbn	99,000
area	95,000
population	86,000

28,000,000 different attribute labels

- Subject Attribute Values

Value	#Rows
usa	135,000
germany	91,000
greece	42,000
new york	59,000
london	37,000
athens	11,000
david beckham	3,000
ronaldinho	1,200
oliver kahn	710
twist shout	2,000
yellow submarine	1,400

1.74 billion rows
253,000,000 different subject labels

HTML-embedded Data

More and more Websites semantically markup the content of their HTML pages.

Microformats

RDFa

Microdata

Schema.org

Google

YAHOO!

Microsoft

schema.org Search

Home Schemas Documentation

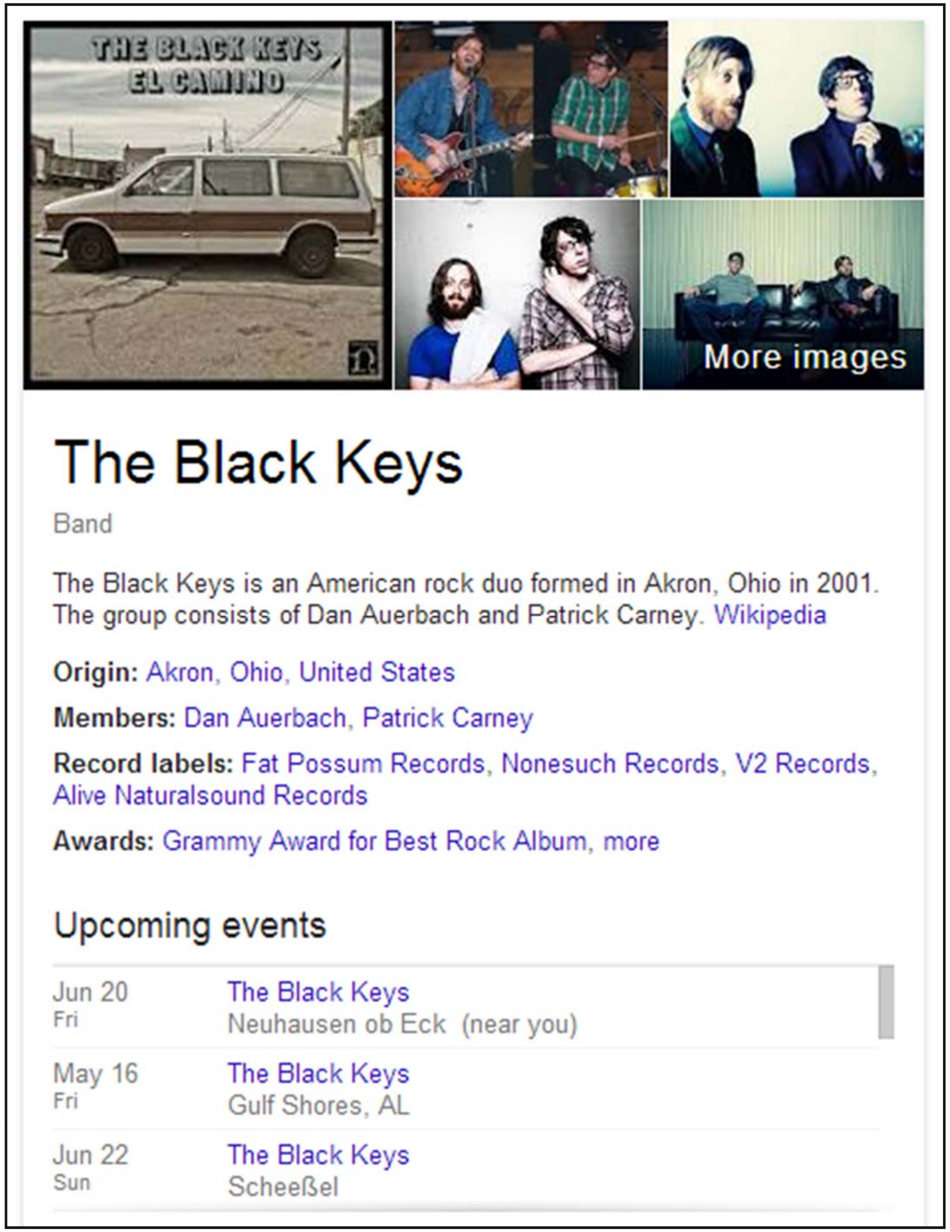
Thing > Organization > LocalBusiness

A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

Property	Expected Type	Description
Properties from Thing		
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Place		
address	PostalAddress	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
containedIn	Place	The basic containment relation between places.

- ask site owners to embed data to enrich search results.
- 200+ Classes: Product, Review, LocalBusiness, Person, Place, Event, …
- Encoding: Microdata or RDFa

Usage of Schema.org Data @ Google



Data snippets within search results

Data snippets within info boxes

Websites containing Structured Data (2012)

Web Data Commons - Microformat, Microdata, RDFa Corpus

- 7 billion RDF triples from Common Crawl 2012
- Winter 2013 release upcoming

369 million of the 3 billion pages contain Microformat, Microdata or RDFa data (12.3%).

2.29 million websites (PLDs) out of 40.6 million provide Microformat, Microdata or RDFa data (5.65%)

Google, October 2013:
15% of all websites provide structured data.

Top Classes Microdata (2012)

	Class	PLDs Total #	PLDs Total %	PLDs in Alexa #	PLDs in Alexa %
1	*schema:BlogPosting*	25,235	17.98	1,502	6.63
2	*datavoc:Breadcrumb*	21,729	15.49	5,244	23.13
3	*schema:PostalAddress*	19,592	13.96	1,404	6.19
4	*schema:Product*	16,612	11.84	3,038	13.40
5	*schema:LocalBusiness*	16,383	11.68	845	3.73
6	*schema:Article*	15,718	11.20	3,025	13.35
7	*datavoc:Review-aggregate*	8,517	6.07	2,376	10.48
8	*schema:Offer*	8,456	6.03	1,474	6.50
9	*datavoc:Rating*	7,711	5.50	1,726	7.61
10	*schema:AggregateRating*	7,029	5.01	1,791	7.90
11	*schema:Organization*	7,011	5.00	1,270	5.60
12	*datavoc:Product*	6,770	4.82	1,156	5.10
13	*schema:WebPage*	6,678	4.76	2,112	9.32
14	*datavoc:Organization*	5,853	4.17	654	2.89
15	*datavoc:Address*	5,559	3.96	654	2.89
16	*schema:Person*	5,237	3.73	890	3.93
17	*schema:GeoCoordinates*	4,677	3.33	312	1.38
18	*schema:Place*	4,131	2.94	488	2.15
19	*schema:Event*	4,102	2.92	659	2.91
20	*datavoc:Person*	2,877	2.05	523	2.31
21	*datavoc:Review*	2,816	2.01	783	3.45

- schema = Schema.org
- datavoc = Google's Rich Snippet Vocabulary

Example: Microdata Local Business

```
<div itemscope itemtype="http://schema.org/LocalBusiness">
  <h1><span itemprop="name">Beachwalk Beachwear & Giftware</span></h1>
  <span itemprop="description"> A superb collection of fine gifts and clothing
  to accent your stay in Mexico Beach.</span>
  <div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
    <span itemprop="streetAddress">3102 Highway 98</span>
    <span itemprop="addressLocality">Mexico Beach</span>,
    <span itemprop="addressRegion">FL</span>
  </div>
  Phone: <span itemprop="telephone">850-648-4200</span>
</div>
```

Looking Deeper into the E-Commerce Data

Microdata Product (2012)

	Property	PLDs Total #	PLDs Total %
1	*dc:title*	16,488	99.25
2	*schema:Product/name*	14,342	86.34
3	*schema:Product/description*	10,297	61.99
4	*schema:Product/image*	8,093	48.72
5	*schema:Product/offers*	7,545	45.42
6	*schema:Offer/price*	6,894	41.50
7	*schema:AggregateRating*	4,308	25.93
8	*schema:AggregateRating/ratingValue*	3,990	24.02
9	*schema:PostalAddress/streetAddress*	3,723	22.41
10	*schema:PostalAddress/addressRegion*	3,502	21.08

Example Name:

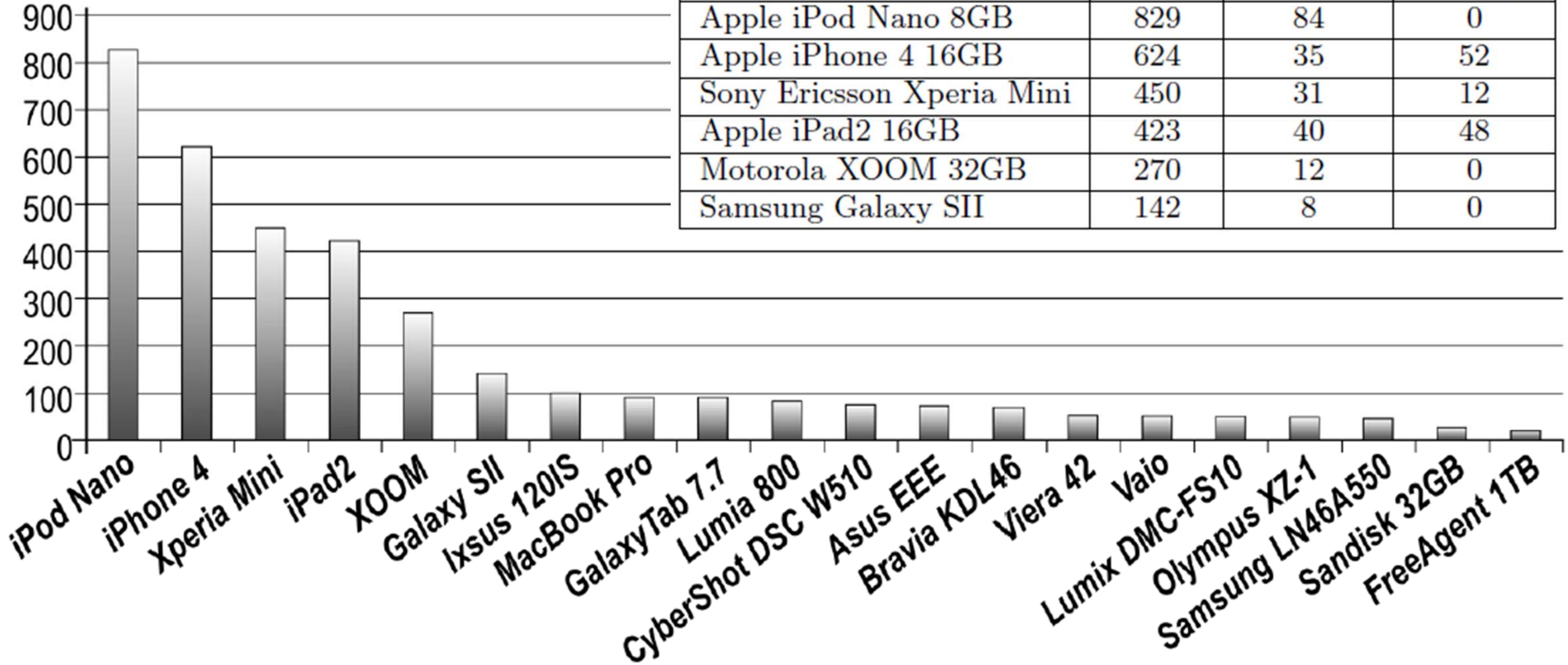
- Apple MacBook Air 11-in, Intel Core i5 1.60GHz, 64 GB, Lion 10.7

Example Description:

- Configured with Intel Core 2 Duo processor, faster Flash Storage with 64 GB Solid State Drive and USB 3.0 …

Identity Resolution for Electronic Products

- We trained parser for product descriptions on data from Amazon.
- We analyzed 1.9 million product offers from 9200 e-shops

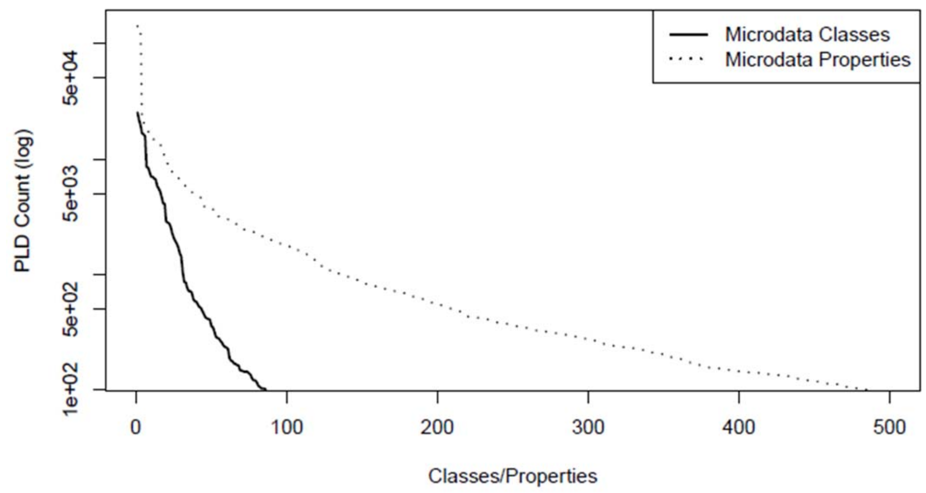


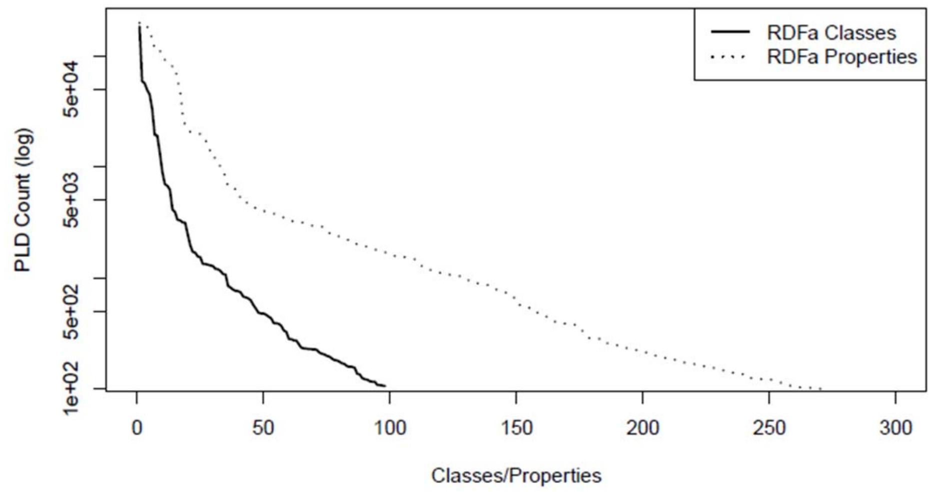
Product	Offers	Reviews	Ratings
Apple iPod Nano 8GB	829	84	0
Apple iPhone 4 16GB	624	35	52
Sony Ericsson Xperia Mini	450	31	12
Apple iPad2 16GB	423	40	48
Motorola XOOM 32GB	270	12	0
Samsung Galaxy SII	142	8	0

- Petar Petrovski, et al.: Integrating Product Data from Websites offering Microdata Markup. DEOS workshop @ WWW 2014.

Representing HTML-embedded Data in Tabular Form

- **Table Generation**
  - Represent each class per website as separate table.
  - subject column: naming convention itemprop="name"
- **Resulting tables**
  - several million tables
  - mostly 2-10 attributes wide
  - up to 100.000s of rows
- **Heterogeneity**
  - low as data providers use vocabularies recommended by Google, Microsoft, Yahoo, and Facebook

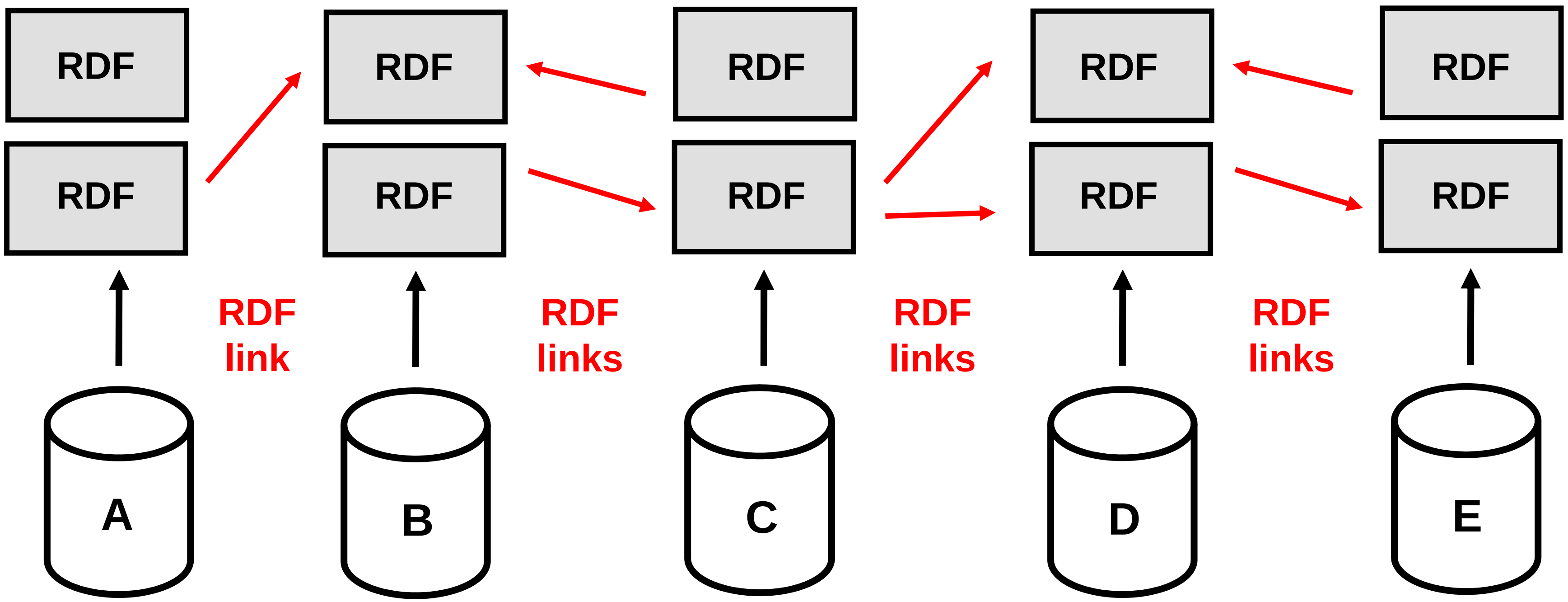




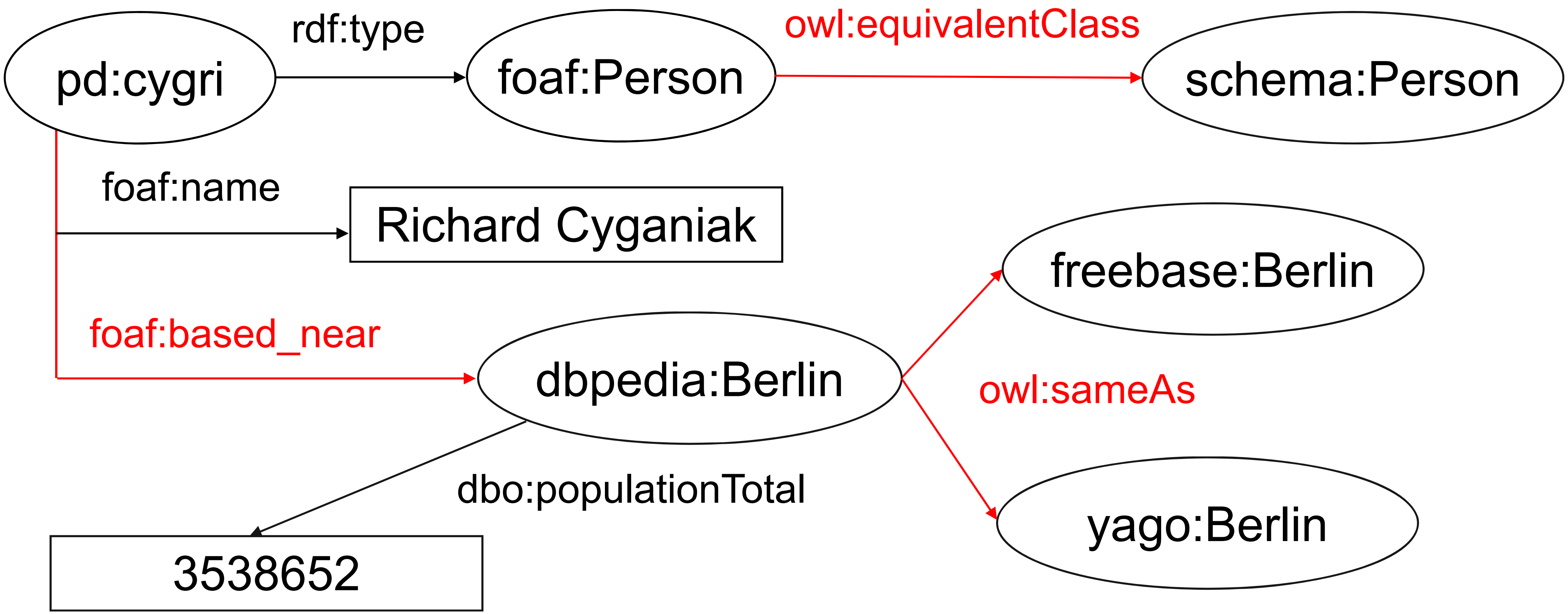
- Bizer, et al.: Deployment of RDFa, Microdata, and Microformats on the Web. ISWC 2013.

Extends the Web with a single global data graph

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



Data Graph including Integration Hints



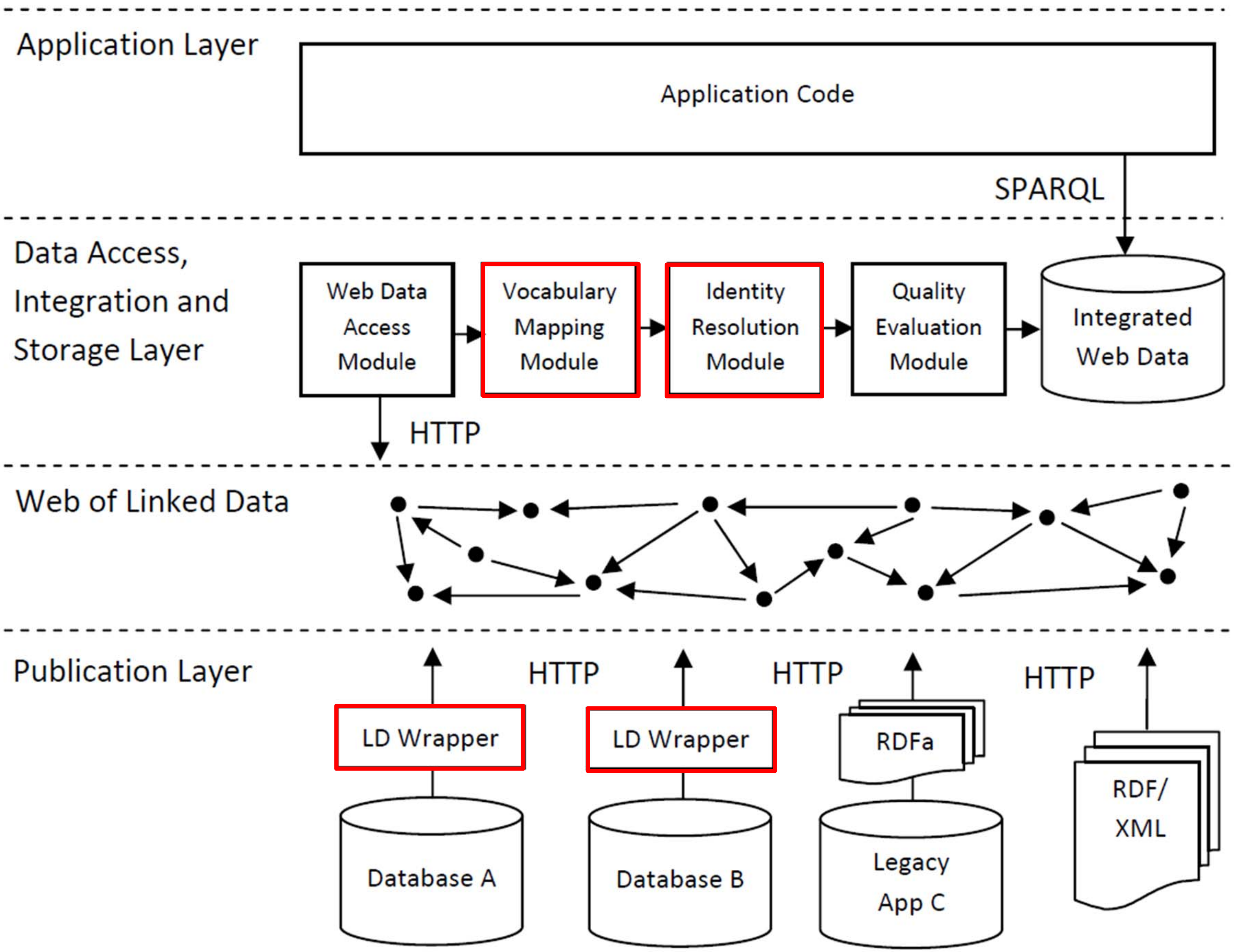
- Data Providers
  - set instance-level and schema-level RDF links
  - reuse terms from common vocabularies

Effort Distribution between Publisher and Consumer

Consumer calculates links and correspondences

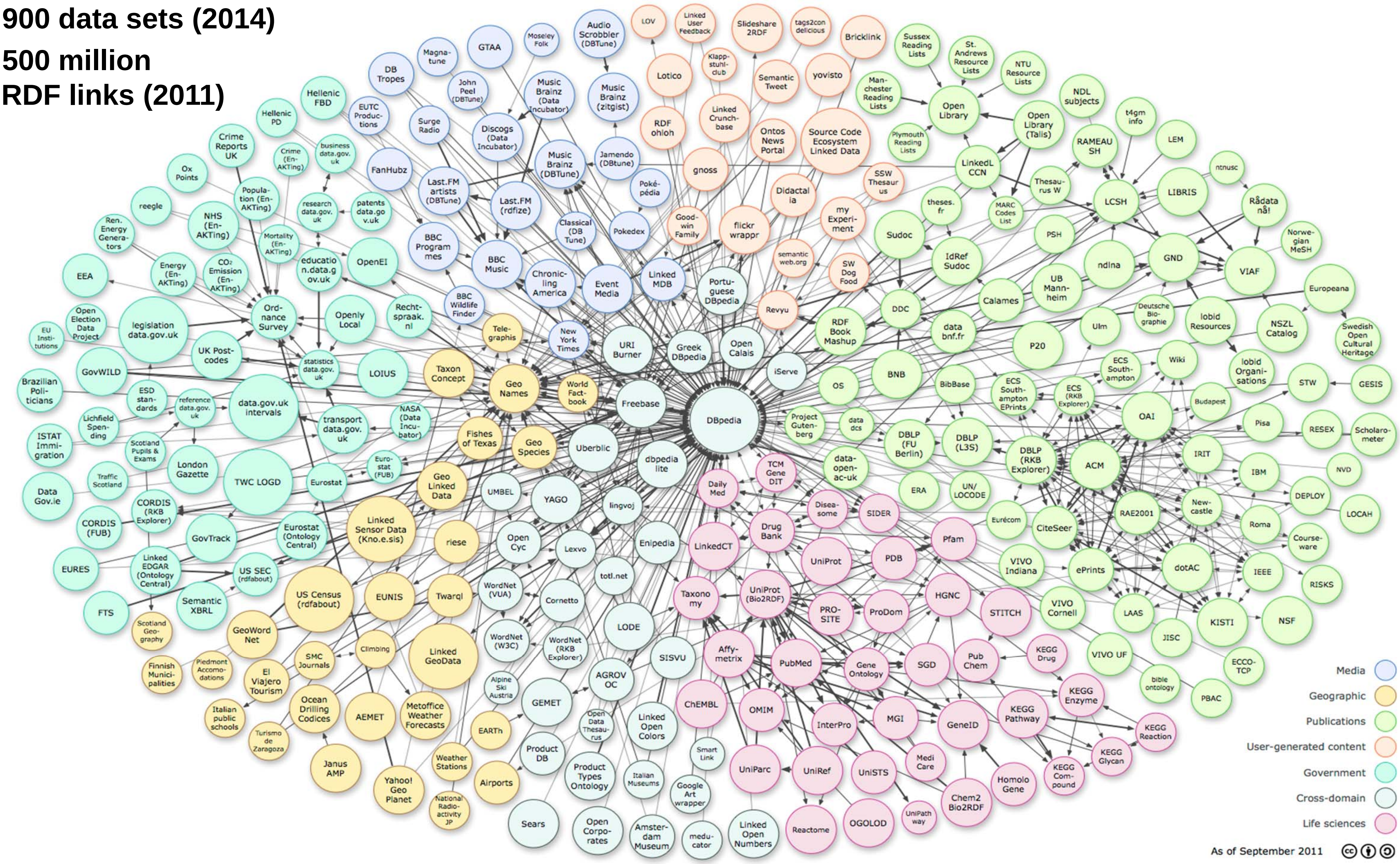
Publisher reuses vocabularies

Publisher or third party publishes links and correspondences



Linked Data on the Web

- ~ **900 data sets (2014)**
- ~ **500 million RDF links (2011)**



Wikipedia Data available as Linked Data

Wikipedia
The Free Encyclopedia

DBpedia

yago
select knowledge

Freebase

Universität Mannheim

Universität Leipzig

max planck institut
informatik

Google

Representing Linked Data in Tabular Form

- Most Linked Data sources have a rather regular structure
  - as they are generated from relational databases
- Table generation
  - generate one table per class and data source
  - use RDF property labels as attribute names
  - subject attribute: naming convention rdfs:label, foaf:name
- Representation of RDF Links
  - Add ID attribute containing original URI of each entity
  - For each link predicate type add two attributes
    1. URI reference to target entity
    2. rdfs:label of target entity

Profile of the Resulting Tables

- Billion Triples Challenge 2012 Dataset
  - 53,000 tables

	Min	Max	Avg
Attributes	3	1,479	**9**
Data Rows	1	372,000	**180**

- DBpedia as Tables
  - 365 tables

	Min	Max	Avg
Attributes	7	730	**498**
Data Rows	1	577,000	**12,000**

Heterogeneity

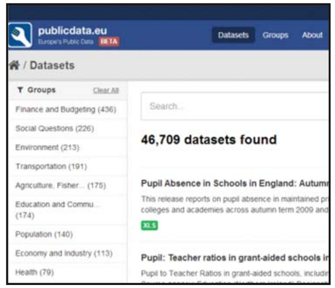
- Low in some domains: People, Publications
- High in other domains: Life Science, eGovernment

- http://km.aifb.kit.edu/projects/btc-2012/
- http://wiki.dbpedia.org/DBpediaAsTables

Data Portals

- datacatalogs.org lists 377 data portals world-wide
  - open government data portals
  - international organizations and NGOs
  - scientific data portals
- Challenges
  - syntax heterogeneity: Excel, CSV, XML, HTML, PDF
  - data values are often time dependent
  - table context understanding and header-unfolding necessary
- Profile of 7600 Tables from PublicData.eu

	Min	Max	Avg	Median
Attributes	2	488	8.88	**9**
Data Rows	1	5,600,000	3,160	**66**



- Ermilov, et al.: User-driven Semantic Mapping of Tabular Data. I-Semantics 2013.

Wrap-up: Structured Data on the Web

- There is lots of data available that
  - we can fit into our data model.
  - use for search join experiments.
- A wide range of topics is covered.
- The size of the tables varies widely.

- Additional types of data sources not considered
  - Web 2.0 APIs, Deep Web via HTML forms
  - HTML Lists, Excel files somewhere on the Web

- Elmeleegy, et al.: Harvesting Relational Tables from Lists on the Web. VLDB-J 2011.
- Chen, Cafarella: Automatic Web Spreadsheet Data Extraction. WS Semantic Search 2013.
- Furche: The Ontological Key: Automatically Understanding and Integrating Forms to access the Deep Web. VLDB-J 2013.

2. Feasibility of Search Joins

Search Join Systems

	Octopus	InfoGather	WikiTables	MSJ Engine
Developer	Google Research University Washington	Microsoft Research Purdue University	Northwestern University	University of Mannheim
Extend Operation	Single Attribute	Single Attribute	Multiple Attributes	Single Attribute Multiple Attributes
Corpus	Google Web crawl via Search API	HTML tables from Bing Web crawl	Tables from Wikipedia	WDC Web Tables Linked Data
Use Case	Data Gathering	Data Gathering	Data Gathering Data Mining	Data Gathering Data Mining

- Cafarella, et al.: Data Integration for the Relational Web. VLDB 2009.
- Yakout, et al.: InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. SIGMOD 2012.
- Bhagavatula, et al.: Methods for Exploring and Mining Tables on Wikipedia. KDD IDEA 2013.

Infogather

- Prototype developed by Microsoft Research
- Operation: Extend with Singe Attribute
- Corpus: 573 million Web tables from Bing Crawl (2011)
- Split HTML Tables into Binary-Entity-Attribute Tables

Subject Attribute Value Attribute

Region	Unemployment
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %

Region	GDP per Capita
Alsace	45.914 €
Lorraine	51.233 €
Guadeloupe	19.810 €

- Yakout, et al.: InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. SIGMOD 2012.

Matching Graph

- Pre-compute matching graph between BEA tables
- Features used for matching:
  1. Attribute label similarity
  2. Attribute values similarity
  3. Key values overlap
  4. Textual context around tables similarity
  5. Table to context similarity
  6. Table to table as bag of words similarity
  7. URL similarity
  8. column width similarity
- Accuracy of the resulting correspondences:
  - Cameras and movies: 0.95
  - Governors and members of parliament: 0.5

Query Processing

- Input: Query table, extension attribute
- Web tables considered relevant for query
  - share subject attribute value with query table
  - and directly match extension attribute
  - or are connected to directly matching tables via matching graph
- Matching score
  - directly matching tables: entity overlap / min($|t_q|$, $|t_w|$)
  - indirectly matching tables: propagate score along edges of matching graph
- Predict values
  - cluster by value
  - sum matching scores per cluster
  - choose centroid of cluster with highest score or top-k centroids

Experimental Setting

- Queries

Query	Subject Attribute	Extension Attribute
Cameras	Camera model	Brand
Movies	Movie name	Director
Baseball	Team name	Player
Albums	Musical band	Album
UK-pm	UK political party	Member of parliament
US-gov	US state	Governor

- Ground Truth
  - Camera, movies: Bing shopping database
  - Baseball, Albums, UK-pm, US-gov: Wikipedia, Freebase
- Size of query table
  - 12 to 6000 rows

Evaluation Results

Query	Subject attribute	Extension Attribute	Precision	Coverage
Cameras	Camera model	Brand	0.85	0.93
Movies	Movie name	Director	0.92	0.97
Baseball	Team name	Player	0.72	1.00
Albums	Musical band	Album	0.75	1.00
UK-pm	UK political party	Member of parliament	0.60	0.91
US-gov	US state	Governor	0.90	1.00
		AVG	**0.79**	**0.97**

Response times: around 100 milliseconds

3. Table Relevance

The Search operator determines the set of relevant Web tables.

$$T_r = s_{q,s,e}(T_{web})$$

Everything on the Web is a claim by somebody.

1. Claims use different surface forms.
   - entity name
   - attribute labels
   - data value
2. Claims refer to a specific point in time.
3. The trustworthiness of claims varies widely.

Dimensions of Table Relevance

1. Entity Coverage
   - Web table should cover many entities in the query table.
2. Attribute Relevance
   - Web table should contain relevant attributes.
3. Timeliness
   - The data should refers to the desired point in time.
4. Trustworthiness
   - The data should be trustworthy.

Dimension: Coverage

Web table should cover many entities of the query table.

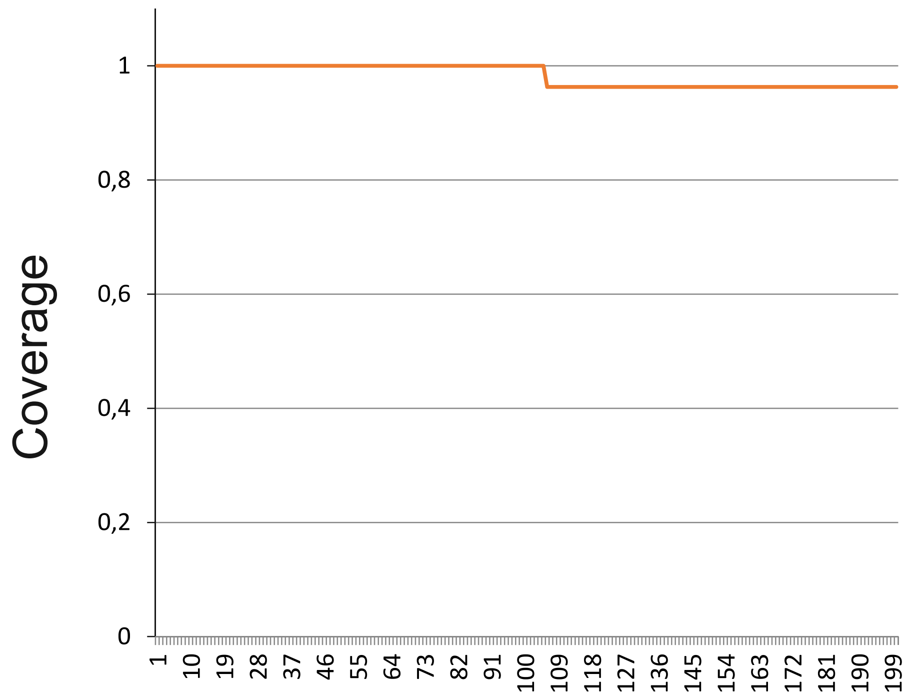
$$c = \frac{entity\ overlap}{|t_q|}$$

- Identity Resolution Approaches
  1. Exact matching on subject attribute
  2. Approximate matching on subject attribute
  3. Matching using external knowledge about surface forms
  4. Matching using multiple attributes from both tables
  5. relying on owl:sameAs links

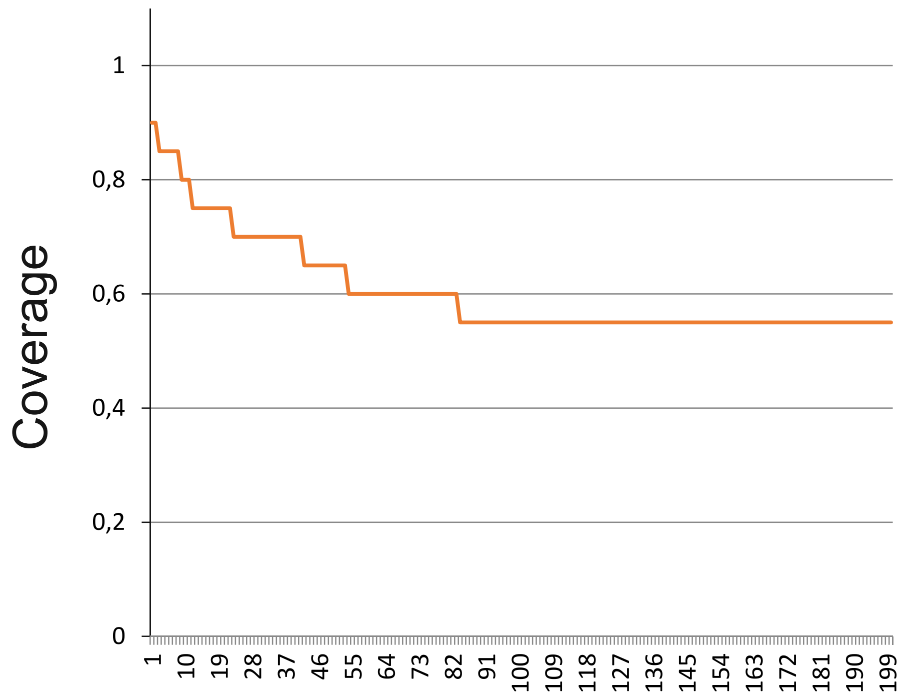
Exact Matching

Exact matching of normalized values against WDC table corpus.

20 country names



15 names of mayor cities



Collective Disambiguation

Query Table

City
Madison
Berlin
Chatham
Perth

Web Table 1

City	Country
Munich	Germany
Berlin	Germany
Mannheim	Germany
Frankfurt	Germany
Karlsruhe	Germany

$$s = \frac{1}{4}$$

Web Table 2

City	Country
Madison	USA
Berlin	USA
Chatham	USA
Fort Reed	USA
Sunville	USA

$$s = \frac{3}{4}$$

Expansion with additional Surface Forms

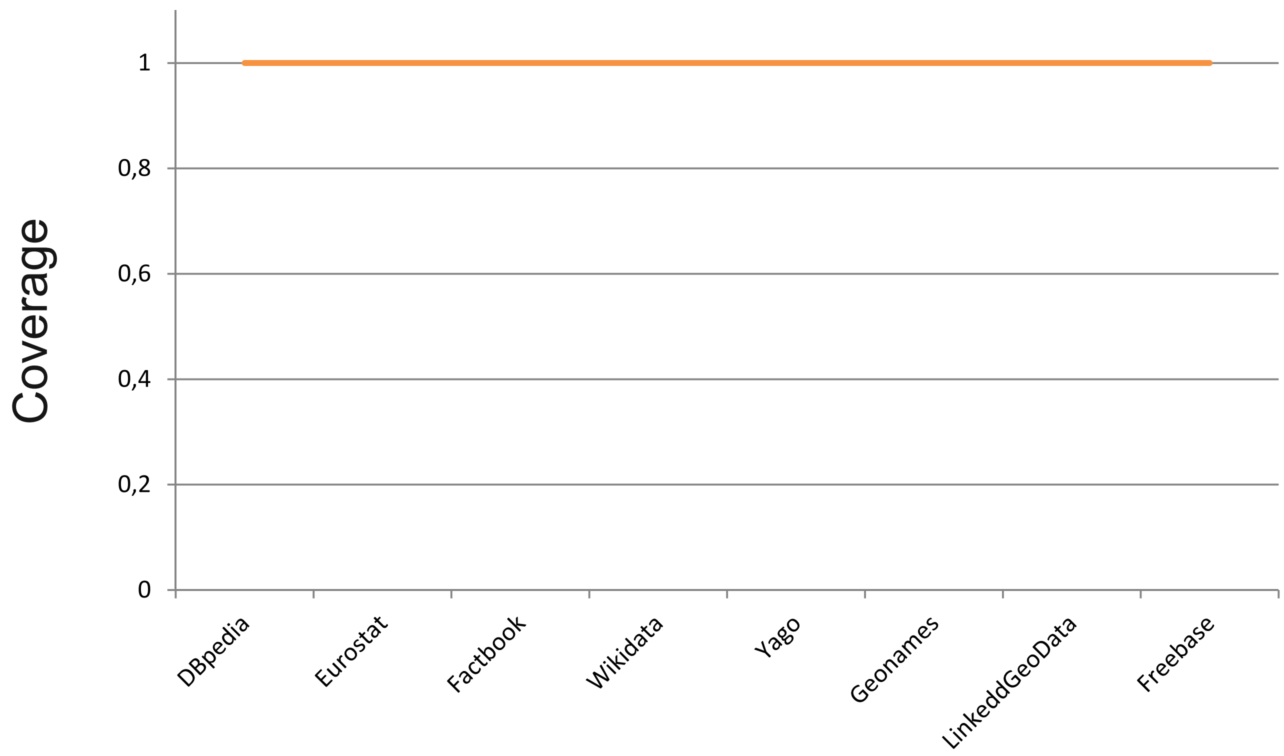
- Examples of surface forms
  - Berlin, 柏林, Berlijn, Berlín, Berlino, Берлин, Berlim, ベルリン
  - FC Bayern München, Bayern Munich, FC Bayern, Bayern München

- Community-generated sources of surface forms
  - Wikipedia
    - Redirects, Cross-Language Links
    - Easy accessible via DBpedia
  - owl:sameAs Links
    - rdfs:labels of interlinked resources
    - 500 million links (2011)

Identity Resolution via owl:sameAs Links

Query table: 20 country names

1. Names are matched to DBpedia
2. owl:sameAs links are followed from DBpedia



Dimension: Attribute Relevance

The Web table should contain relevant attributes.

Attribute relevance depends on query type

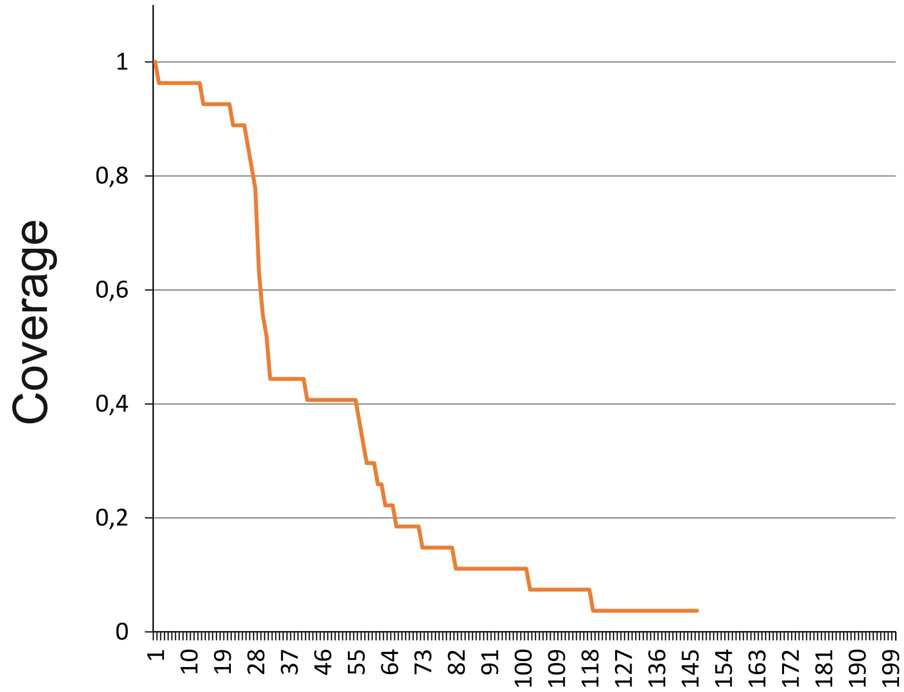
1. Extend with single attribute
   1. tables that have attribute directly matching the keywords
   2. tables that have corresponding attributes
2. Extend with all attributes above density threshold
   1. take all attributes
   2. prefer attributes that are related to attributes in query table

Extend with Single Attribute

Simplest approach: Exact matching of normalized values.

Query table: 20 country names

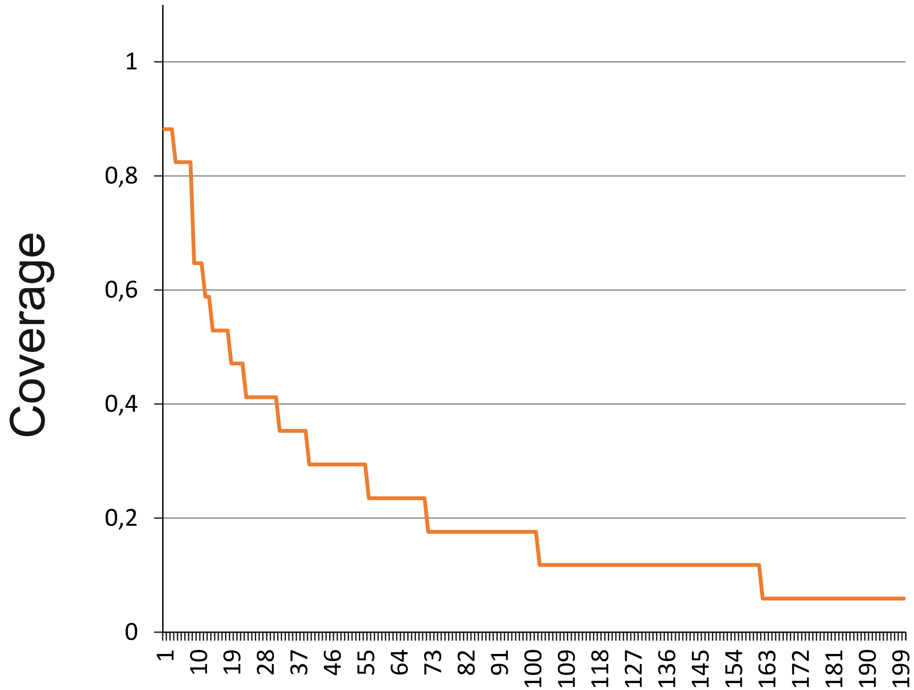
Search attribute: „population“



Median of values differs on average 4% from Wikipedia value.

Query table: 15 names of mayor cities

Search attribute: „country”



Include Corresponding Attributes

- Approaches:

1. attributes having a synonymous name
2. attributes that correspond according to matching
   1. tables with each other
   2. tables against a mediated schema (knowledge base)

- Experimental Results: Yakout et al.
  - Attribute Synonyms
    - Extension attribute value precision: 70 %
  - Attributes corresponding via Schema Matching
    - Extension attribute value precision: 79 %

• Yakout, et al.: InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. SIGMOD 2012.

Mediated Schemata

- Cross-domain knowledge bases
  - **DBpedia**
    - Classes: 259; attributes 1,373; entities: 4 million
  - **Freebase**
    - Classes 1,450; attributes 3,500; entities: 15 million
- Advantage
  - matching against large KBs is likely easier than matching small tables
  - Zhang: Web tables against DBpedia matching accuracy: 85%
- Disadvantage
  - Restricted to attributes contained in the knowledge base

- Zhang, et al.: Mapping entity-attribute web tables to web-scale knowledge bases. In: Database Systems for Advanced Applications. Springer, 2013

Probase and Biperpedia

- Build comprehensive KBs using
  - KBs like DBpedia and Freebase as seeds
  - information extraction from Web text
    - taxonomy and instances: Hearst patterns „Y such as X“
    - Attributes: Patterns „What is the A of I?”
  - search engine query logs “New York City inhabitants”
- Probase (Microsoft)
  - 2.7 million classes, a set of attributes for each class
- Biperpedia (Google)
  - 10,000 classes and 67,000 attributes.
  - table annotation accuracy: 51%

- Wu, et al: Probase: a probabilistic taxonomy for text understanding. SIGMOD 2012.
- Wang, et al.: Understanding tables on the Web. ER 2012.
- Ventis, et al.: Recovering Semantics of Tables on the Web. VLDB 2011.
- Gupta, et al.: Biperpedia: An Ontology for Search Applications. VLDB 2014.

Extend Table with Multiple Attributes

Approaches

1. take all attributes
2. prefer attributes that are related to attributes in query table

Take all Attributes

- Query table: 20 country names
- WDC Web tables (Top 200 tables)
  - 920 additional attributes
  - 346 after attribute consolidation
- Linked Data (BTC dataset)
  - 1131 additional attributes
  - 403 after attribute consolidation

- Problem: Number of attributes might overwhelm the user
  (even if he only looks at correlating attributes)

Schema Complement

Add attributes while preserving "consistency" of the schema.

- What should be considered "consistent"?
- Answer: Combinations of local and Web attributes that often occur together on the Web.

- Approach
  - Generate frequent item set database from Web table schema corpus.
  - Retrieve frequency of all combinations of one local and Web attribute.
  - Aggregate frequencies for Web attributes.

- Das Sarma, et al.: Finding related tables. SIGMOD 2012.

Dimension: Timeliness

The data in the Web tables should refers to the desired point in time.

- Requires meta-information about the intended point in time.
- HTML Tables
  - Time reference in table
    - Problem: only present in a few tables
  - Time reference on page
    - Problem: difficult to hard to understand
- Linked Data
  - W3C Data Cube Vocabulary

Region	Unemployment (2013)
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %

The table below provides unemployment data for 2013.
…

Label Propagation

- Approach: Infer time reference by matching to other tables that contain this information.

Region	Unemployment (2013)
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %

2013

Region	Unemployment
Alsace	11.1 %
Lorraine	11.8 %
Guadeloupe	27.3 %
Centre	10.1 %
Martinique	24.7 %

- Evaluation Results:
  - Accuracy: 89%
  - Recall: 50%

- Zhang, Chakrabarti: InfoGather+: Semantic Matching and Annotation of Numeric and Time-varying attributes in Web Tables. SIGMOD 2013.

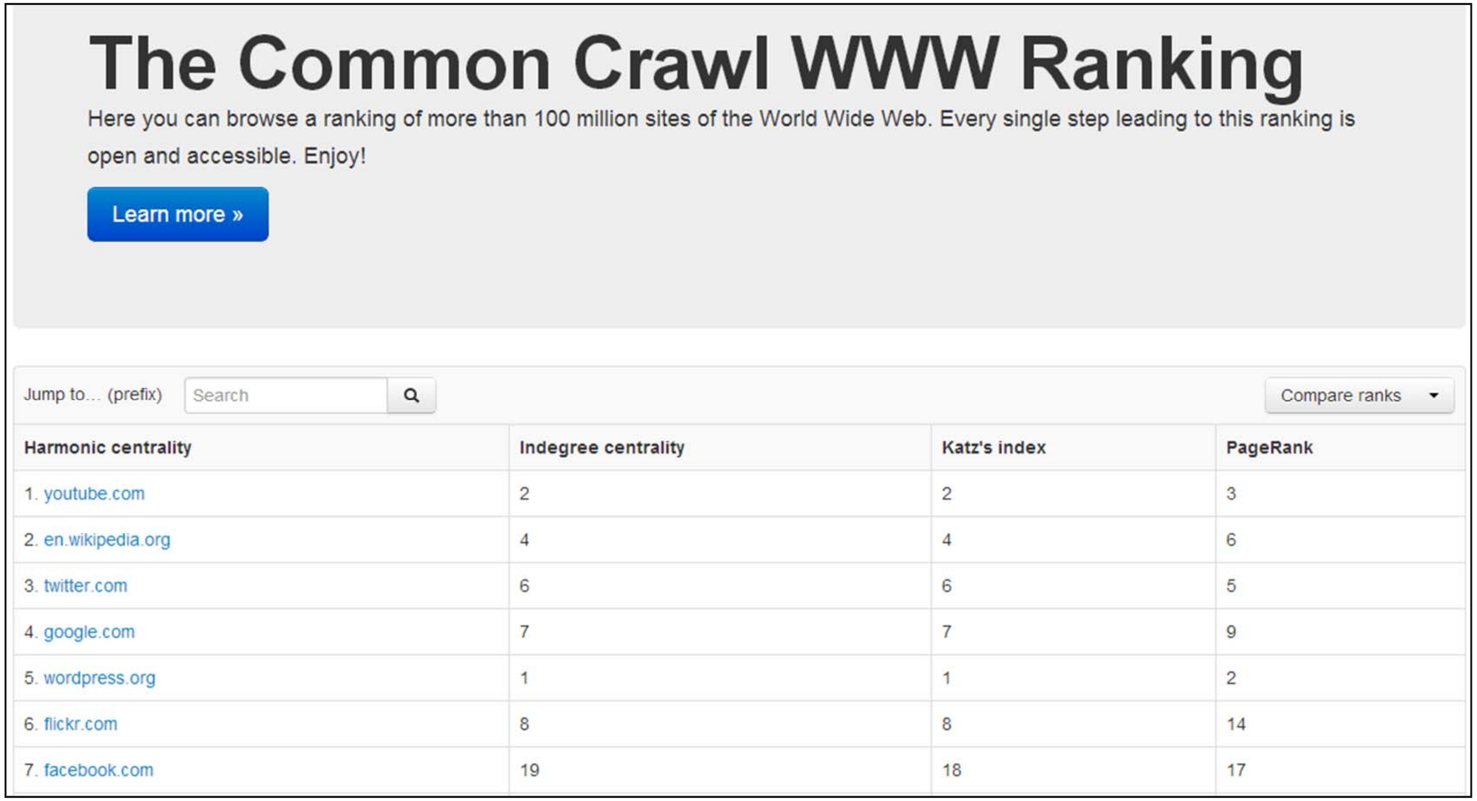
Dimension: Trustworthiness

The data in the Web table should be trustworthy.

- Not considered by Search Join systems yet.
- Approach exploiting only the data itself
  - Consider value consistency between query table and Web table if the tables contain overlapping attributes
- Approaches exploiting external knowledge
  - Put preference on specific sources
    - prefer data from .gov websites
  - Exploit hyperlink structure of the Web

Web Data Commons - Hyperlink Graph

- Covers 3.5 billion web pages and 128 billion hyperlinks
- Extracted from Common Crawl 2012



The Common Crawl WWW Ranking

Here you can browse a ranking of more than 100 million sites of the World Wide Web. Every single step leading to this ranking is open and accessible. Enjoy!

Learn more »

Jump to… (prefix) Search

Compare ranks

Harmonic centrality	Indegree centrality	Katz's index	PageRank
1. youtube.com	2	2	3
2. en.wikipedia.org	4	4	6
3. twitter.com	6	6	5
4. google.com	7	7	9
5. wordpress.org	1	1	2
6. flickr.com	8	8	14
7. facebook.com	19	18	17

- http://webdatacommons.org/hyperlinkgraph/
- http://wwwranking.webdatacommons.org/

Wrap-up: Dimensions of Table Relevance

1. Entity Coverage
   – Web table should cover many entities in the query table.
2. Attribute Relevance
   – Web table should contain relevant attributes.
3. Timeliness
   – The data should refers to the desired point in time.
4. Trustworthiness
   – The data should be trustworthy.

Conclusion

Search Joins bring together Web Search and DB Joins via the concept of table relevance.

- Simple queries are feasible with “acceptable”(?) precision.
- The Web is one application domain for search joins, corporate intranets are the other.