

Statistical Natural Language Processing Colloquium
Universität Heidelberg, 26.11.2015

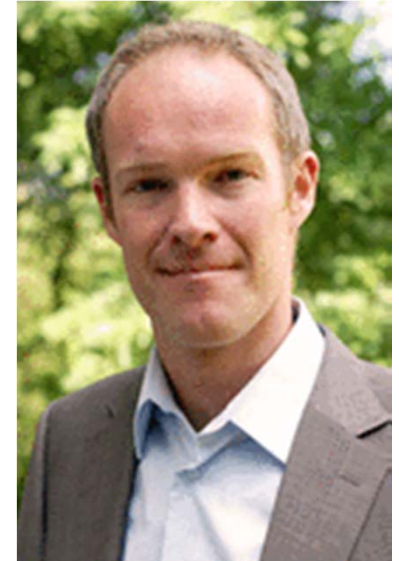
Data Search and Search Joins

Prof. Dr. Christian Bizer

Hello

Professor Christian Bizer

University of Mannheim



Research Topics

- Web Technologies
- Web Data Profiling
- Web Data Integration
- Web Mining

1. Research methods for integrating and mining large amounts of heterogeneous information from the Web.
2. Empirically analyze the content and structure of the Web.

■ 5 Professors

- Heiner Stuckenschmidt
- Rainer Gemulla
- Christian Bizer
- Simone Ponzetto
- Heiko Paulheim



■ 25 postdocs and PhD students

■ <http://dws.informatik.uni-mannheim.de/>

1. Motivation for Data Search
2. Types of Data Search
 1. Entity Search
 2. Table Search
 3. Constraint and Unconstraint Search Joins
3. The Mannheim Search Join Engine
 1. Methods
 2. Evaluation
4. Conclusion and Outlook

1. Motivation for Data Search

Deluge of structured data on the Web

1. Semantic annotations in HTML pages
2. Linked data
3. Relational HTML tables
4. Data portals

Need for search techniques that exploit the structure of the data.

Semantic Annotations in HTML Pages

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



- ask site owners to embed schema.org annotations into their HTML pages
- 200+ Classes: Product, Review, LocalBusiness, Person, Place, Event, ...
- Encoding: Microdata or RDFa

The screenshot shows the schema.org website. At the top is a red header with the 'schema.org' logo, a search bar, and a 'Search' button. Below the header is a navigation bar with links for 'Home', 'Schemas', and 'Documentation'. The main content area displays the breadcrumb path 'Thing > Organization > LocalBusiness'. Below this path is a descriptive paragraph: '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.' Below the text is a table with three columns: 'Property', 'Expected Type', and 'Description'. The table is divided into two sections: 'Properties from Thing' and 'Properties from Place'.

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.

Usage of Schema.org Data @ Google

Gramercy Tavern - Flatiron - New York, NY | Yelp

www.yelp.com > Restaurants > American (New) ▾

★★★★★ Rating: 4.5 - 1,288 reviews - Price range: \$\$\$\$

Jeff C and I were in New York for vacation, and I wanted to treat him to a nice dinner for Gramercy Tavern is certainly a legendary NY dining establishment.

Gramercy Tavern Restaurant - New York, NY | OpenTable

www.opentable.com > ... > Gramercy restaurants ▾

★★★★★ Rating: 4.7 - 508 reviews - Price range: \$50 and over

Book now at Gramercy Tavern in New York, explore menu, see photos and read 508 reviews: "The menu was so limited but it was worth trying, food was deli..."

Data snippets
within
search results

Data snippets
within
info boxes



The Black Keys

Band

The Black Keys is an American rock duo formed in Akron, Ohio in 2001. The group consists of Dan Auerbach and Patrick Carney. [Wikipedia](#)

Origin: [Akron, Ohio, United States](#)

Members: [Dan Auerbach](#), [Patrick Carney](#)

Record labels: [Fat Possum Records](#), [Nonesuch Records](#), [V2 Records](#), [Alive Natural Sound Records](#)

Awards: [Grammy Award for Best Rock Album](#), [more](#)

Upcoming events

Jun 20 Fri	The Black Keys Neuhausen ob Eck (near you)
May 16 Fri	The Black Keys Gulf Shores, AL
Jun 22 Sun	The Black Keys Scheeßel

Adoption of Semantic Annotations in 2014

WebDataCommons project extracted all Microformat, Microdata, RDFa data from the Common Crawl 2014

620 million HTML pages out of the 2 billion pages provide semantic annotations (30%).

2.72 million pay-level-domains (PLDs) out of the 15.68 million pay-level-domains covered by the crawl provide annotations (17%).

**Google, 2014*:
5 million websites provide Schema.org data.**

* Guha in LDOW2014 Keynote

Topical Focus – Microdata 2014

■ Top Classes

■ Topics:

- CMS and blog metadata
- products and offers
- ratings and reviews
- business listings
- address data
- ...and a massive long tail

	Class	Instances #	2014		2013	
			PLDs		PLDs	
			#	%	#	%
1	schema:WebPage	51.757.000	148,893	18,16%	69.712	15,04
2	schema:Article	54.972.000	88,7	10,82%	65.930	14,22
3	schema:Blog	3.787.000	110,663	13,50%	64.709	13,96
4	schema:Product	288.083.000	89,608	10,93%	56.388	12,16
5	schema:PostalAddress	48.804.000	101,086	12,33%	52.446	11,31
6	dv:Breadcrumb	269.088.000	76,894	9,38%	44.187	9,53
7	schema:AggregateRating	59.070.000	50,510	6,16%	36.823	7,94
8	schema:Offer	236.953.000	62,849	7,66%	35.635	7,69
9	schema:LocalBusiness	20.194.000	62,191	7,58%	35.264	7,61
10	schema:BlogPosting	11.458.000	65,397	7,98%	32.056	6,92
11	schema:Organization	101.769.000	52,733	6,43%	24.255	5,23
12	schema:Person	115.376.000	47,936	5,85%	21.107	4,55
13	schema:ImageObject	35.356.000	25,573	3,12%	16.084	3,47
14	dv:Product	12.411.000	16,003	1,95%	13.844	2,99
15	schema:Review	42.561.000	20,124	2,45%	13.137	2,83
16	dv:Review-aggregate	3.964.000	14,094	1,72%	13.075	2,82
17	dv:Organization	3.155.000	10,649	1,30%	9.582	2,07
18	dv:Offer	7.170.000	11,64	1,42%	9.298	2,01
19	dv:Address	2.138.000	9,674	1,18%	8.866	1,91
20	dv:Rating	1.732.000	9,367	1,14%	8.360	1,8

schema: = Schema.org

dv: = Google Rich Snippet Vocabulary (deprecated)

Adoption by E-Commerce Websites

Distribution by Top-Level Domain

TLD	#PLDs
com	38344
co.uk	3605
net	1813
de	1333
pl	1273
com.br	1194
ru	1165
com.au	1062
nl	1002

Adoption by Top-15:
60 %

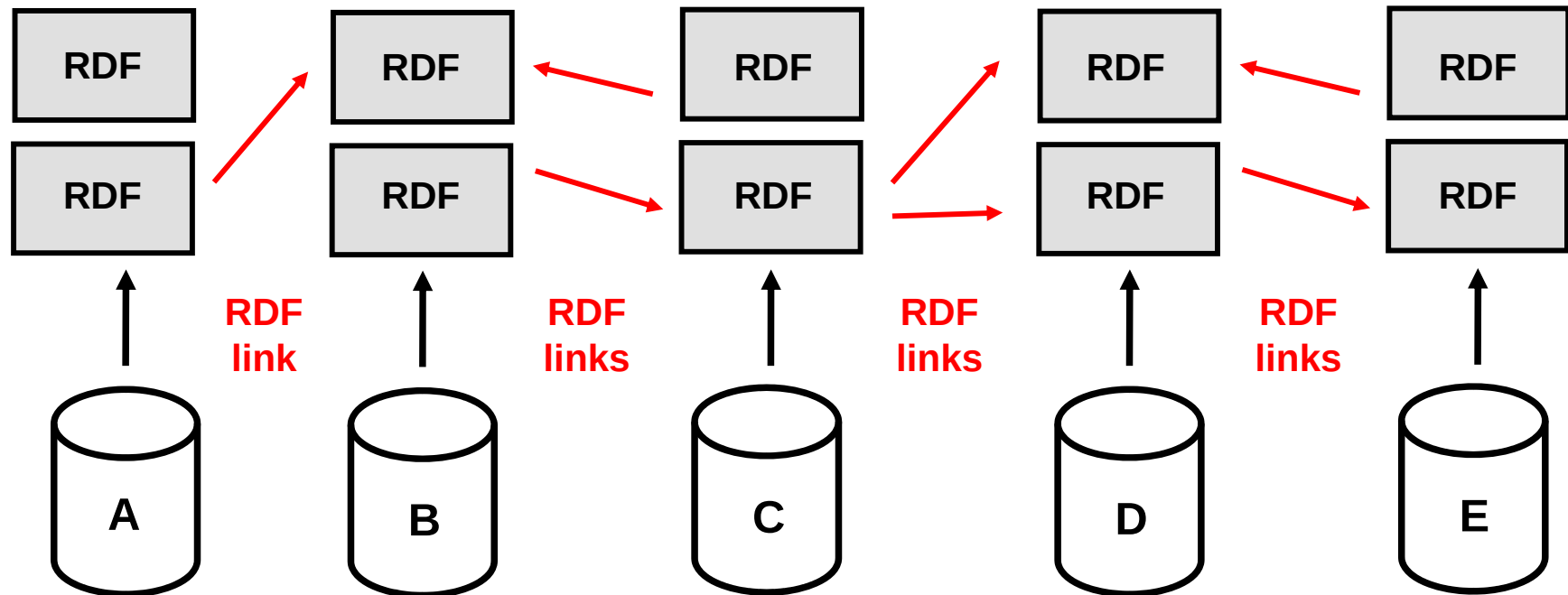
Alexa Top-15 Shopping Sites

Website	schema:Product
Amazon.com	☒
Ebay.com	✓
NetFlix.com	☒
Amazon.co.uk	☒
Walmart.com	✓
etsy.com	☒
Ikea.com	✓
Bestbuy.com	✓
Homedepot.com	✓
Target.com	✓
Groupon.com	☒
Newegg.com	✓
Lowes.com	☒
Macys.com	✓
Nordstrom.com	✓

Linked Data

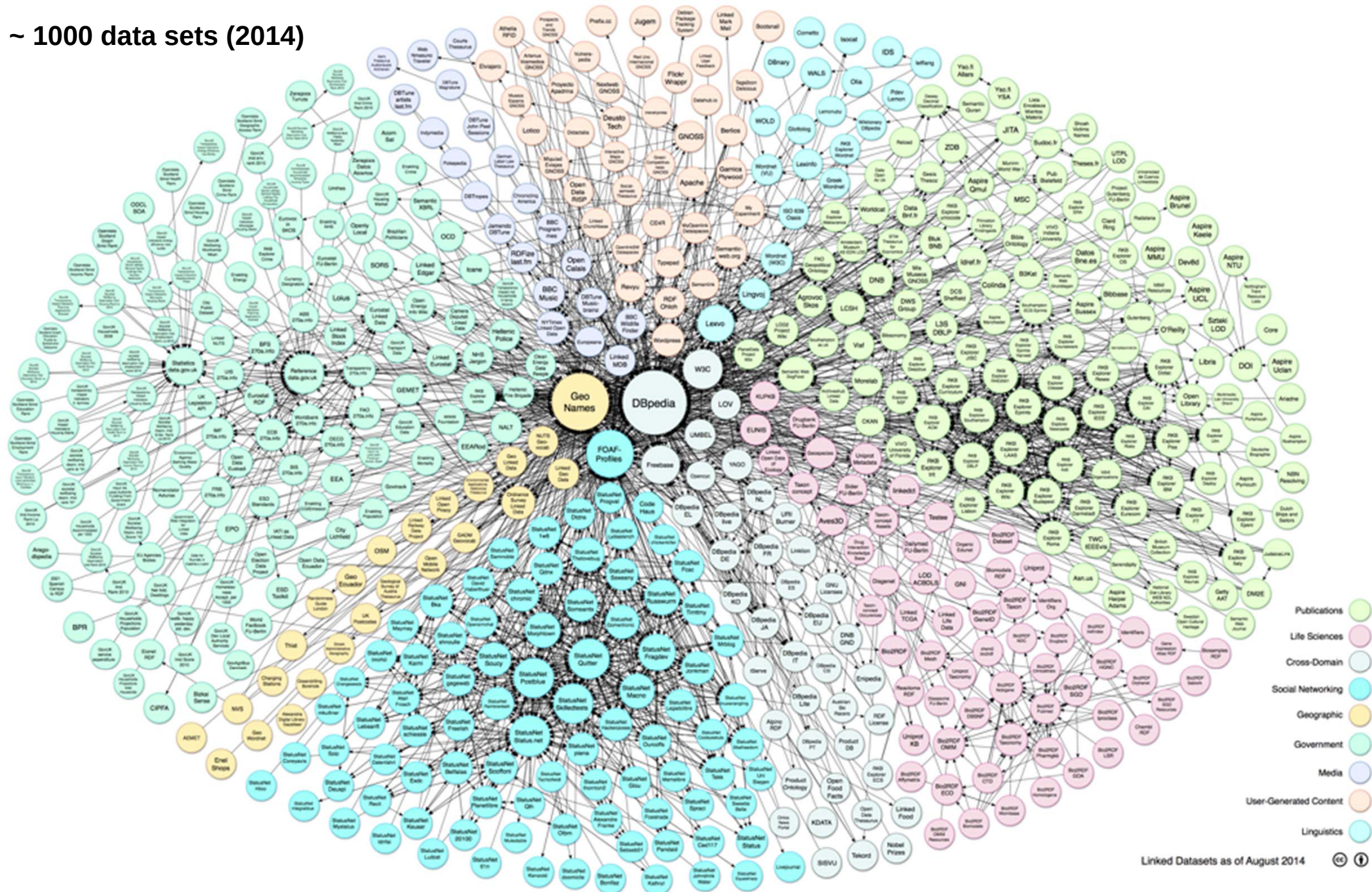
Extends the Web with a **single global data graph**

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



Linked Data on the Web

~ 1000 data sets (2014)



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	

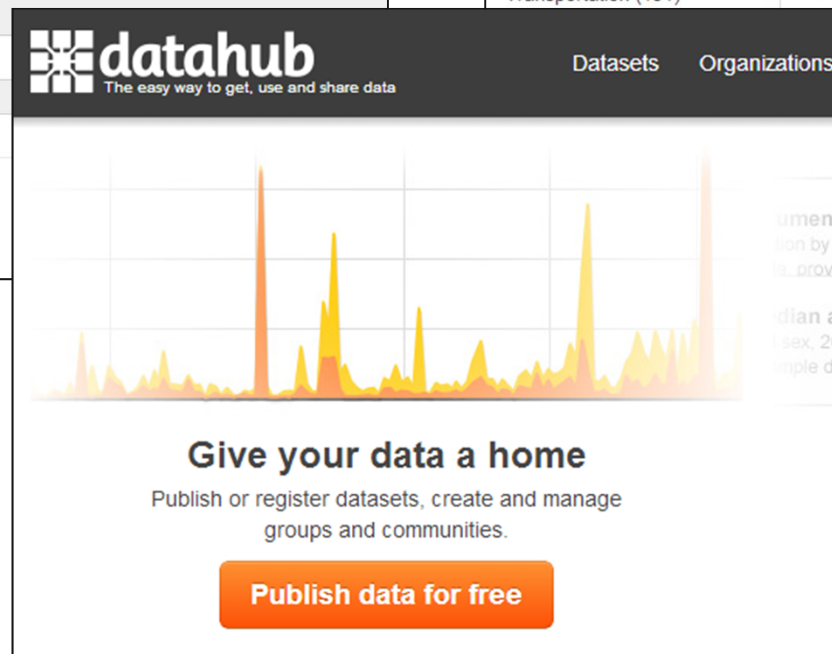
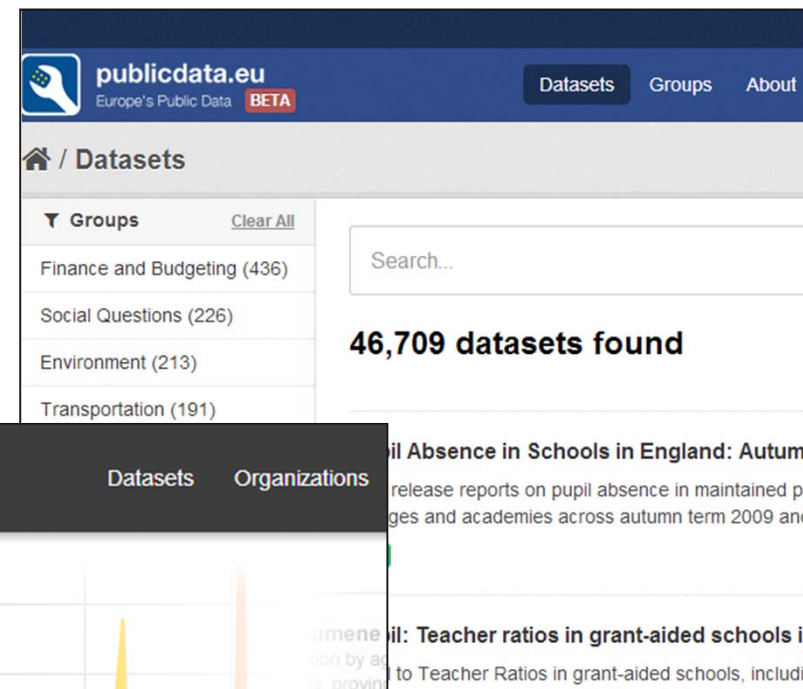
Largest German cities			
German cities ranked 1 to 100 German cities ranked 101 to 200 German cities ranked 201 to 300			
Rank	City	State	Population
1	Berlin	Berlin	3,275,000
2	Hamburg	Hamburg	1,686,100
3	München	Bavaria	1,185,400
4	Köln	Northrhine-Westfalia	965,300
5	Frankfurt	Hessen	648,000
		Northrhine-Westfalia	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.
- WebDataCommons Web Tabel Corpus 2015: 233 million tables for public download.

Data Portals

Several 100.000 datasets are available via data portals.



Summary: Motivation for Data Search

Deluge of structured data on the Web

1. Semantic annotations in HTML pages
2. Linked data
3. Relational HTML tables
4. Data portals

Deluge of structured data on the Intranet

1. Piles of Excel files
2. Piles of CSV files
3. Other data dumps

2. Types of Data Search

1. Entity Search
2. Table Search
3. Constraint and Unconstraint Search Joins

Entity Search

Given some keywords and optionally some facet filters, generate ranked list of **relevant entities**.

- Well researched area with mass market adoption

The screenshot shows the Falcon search engine interface. The search bar contains 'Berlin'. On the left, there is a sidebar with a 'Type' section listing various entity types like 'Any type', 'affelding', 'agent', etc. The main content area displays 'Objects 1 - 5 of 100 for your search Berlin'. The first result is 'Berlin - City, Subject, States of Germany' with detailed metadata including type, label, homepage, and wiki page sort key. The second result is 'Ost-Berlin - Thing, Place, Place' with a comment and subject information.

The screenshot shows a Google Shopping search for 'samsung galaxy s6'. The search bar at the top contains 'samsung galaxy s6'. Below the search bar, there are tabs for 'Web', 'Images', 'Maps', 'Shopping', and 'More'. The 'Shopping' tab is selected. The results show two product listings. The first listing is 'Samsung Galaxy S6 - 32 GB - Black Sapphire - T-Mobile - GSM' with a price of \$489.99 and 3,092 product reviews. The second listing is 'Samsung Galaxy S6 G920 32GB Factory Unlocked Cell Phone for GS...' with a price of \$499.99 and 13 product reviews. Both listings include images of the phone and detailed specifications.

■ Techniques

- hits on different fields contribute differently to ranking
- surface forms of entity names are gathered via click log analysis
- query refinement using facets

Table Search

Given some keywords generate ranked list of **relevant tables**.

■ Example: Google Table Search

The screenshot shows the Google Table Search interface. At the top, the Google logo is on the left, and a search bar contains the text 'Universities in Germany' with a magnifying glass icon on the right. Below the search bar, the word 'Tables' is followed by 'experimental' in smaller text. To the right, it says 'Ergebnisse 1 - 10 von ungefähr 178.971 für Universities in Germany (0.41 Sekunden)'. On the left side, there is a sidebar with links: 'Web', 'Web Tables', 'Fusion Tables', and 'Send Feedback'. The main content area displays two search results. The first result is titled 'List of universities in Germany - Wikipedia, the free encyclopedia' with the URL 'https://en.wikipedia.org/wiki/List_of_universities_in_Germany'. Below the title, there is a table snippet with columns 'University', 'University of ...', 'Leipzig University', and 'University of ...'. Below the snippet, it says 'Show less (11 rows / 4 columns total) - Export data'. There are two buttons: 'Export to Google Sheets' and 'Export to FusionTables'. Below these buttons is a table with 4 columns: 'University', 'Year of establishment', 'Students', and 'Location'. The table contains 11 rows of data. The second result is titled 'List of Universities in Germany - Study in Germany for Free' with the URL 'http://www.studying-in-germany.org/list-of-universities-in-germany/'. Below the title, there is a table snippet with columns 'University', 'Technische Universität ...', 'Otto Beisheim ...', and 'Universität ...'. Below the snippet, it says 'Show more (21 rows / 3 columns total) - Export data'.

Google Universities in Germany

Tables experimental Ergebnisse 1 - 10 von ungefähr 178.971 für Universities in Germany (0.41 Sekunden)

Web

Web Tables

Fusion Tables

Send Feedback

[List of universities in Germany - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/List_of_universities_in_Germany)
https://en.wikipedia.org/wiki/List_of_universities_in_Germany

University	University of ...	Leipzig University	University of ...
University of Heidelberg	1386	31,500	Heidelberg
Leipzig University	1409	29,500	Leipzig
University of Rostock	1419	14,000	Rostock
University of Greifswald	1456	12,000	Greifswald
University of Freiburg	1457	24,700	Freiburg
University of Munich	1472	44,000	Munich
University of Tübingen	1477	28,700	Tübingen
University of Halle-Wittenberg	1502	18,500	Halle
University of Marburg	1527	24,000	Marburg
University of Jena	1558	22,000	Jena

Show less (11 rows / 4 columns total) - Export data

Export to Google Sheets Export to FusionTables

[List of Universities in Germany - Study in Germany for Free](http://www.studying-in-germany.org/list-of-universities-in-germany/)
<http://www.studying-in-germany.org/list-of-universities-in-germany/>

University	Technische Universität ...	Otto Beisheim ...	Universität ...
------------	----------------------------	-------------------	-----------------

Show more (21 rows / 3 columns total) - Export data

Table Search: FeatureRank

- Combination of query independent and query dependent features
- Weights learned using linear regression
- Heavily weighted features
 - hits left-most column
 - hits header
- Result quality: fraction of high scoring relevant tables

rows
cols
has-header?
of NULLs in table
document-search rank of source page
hits on header
hits on leftmost column
hits on second-to-leftmost column
hits on table body

k	Naïve	FeatureRank
10	0.26	0.43
20	0.33	0.56
30	0.34	0.66

Ranking using Deeper Understanding of Tables

- Improve ranking based on deeper understanding of the tables
- Subject Attribute Detection (Ventis)
 - Simple heuristic approach (Accuracy: 83%)
 - take first column from the left of Web tables 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%
- Match tables against large-scale knowledge base (Ventis)
 - Web IsA database for types, TextRunner for relations

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.

Microsoft Power Query for Excel

- Excel add-in offering table search
- Provides various adapters to index intranet and web data

The screenshot displays the Microsoft Power Query interface. The main window shows a table titled "S&P 500 Component Stocks - List of S&P 500 companies" with columns: Ticker symbol, Company, SEC filings, GICS Sector, and GICS Sub Industry. The table lists various companies like 3M Co., Abbott Laboratories, and AbbVie. Below the table, it shows "Columns [8]", "Last modified Monday, September 16, 2013", "From Wikipedia", and "Data Sources [1]" with the URL "http://en.wikipedia.org/wiki/List_of_S&P_500_companies". At the bottom, there are buttons for "ADD TO WORKSHEET", "EDIT", and "LEGAL TERMS".

The right sidebar, titled "Online Search", shows a search bar with "S&P 500" entered and "All" selected. It indicates "562 results" and lists several search results, including "Recent changes to the list of...", "S&P 500 Component Stocks...", "Investing - S&P 500", "Index maintenance - S&P 500", and "Total annual returns - Mark...".

At the bottom of the interface, there is a status bar showing "30 saugs." and a navigation bar with buttons for "1", "2", "3", "4", "5", and "Next".

Excel files
CVS files
XML files
Text files
SQL databases
Azure marketplace
Sharepoint lists
HDFS
Facebook API
SAP BI Universes
Salesforce Objects

Shortcoming: Manual Data Integration Required

First Trading Day (2) - Query Editor

File **Home** **View**

Query **Reduce** **Sort** **Transform** **Create** **Combine** **Help**

Apply & Refresh Close **Remove Top Rows** **Remove Columns** **Split Column** **Group By** **Replace Values** **Fill Down** **Transform** **Unpivot** **Merge** **Append**

Table.RenameColumns(FilteredRows,{{"NewColumn.exchange", "exchange"}},

exchange	stock_symbol	date	stock_price_close	stock_gain_loss_dollar	stock_gain_loss_p
NYSE	AA	1/2/2009	12.11	0	
NYSE	ABC	1/2/2009	35.57	0	
NYSE	ABT	1/2/2009	53.56	0	
NYSE	ACE	1/2/2009	53.03	0	
NYSE	ACN	1/2/2009	33.67	0	
NYSE	ADI	1/2/2009	19.67	0	
NYSE	ADM	1/2/2009	29.19	0	
NYSE	ADP	1/2/2009	40.41	0	
NYSE	AEE	1/2/2009	34.12	0	
NYSE	AEP	1/2/2009	34	0	
NYSE	AES	1/2/2009	8.64	0	
NYSE	AET	1/2/2009	29.41	0	
NYSE	AFL	1/2/2009	46.28	0	
NYSE	AGN	1/2/2009	41.49	0	
NYSE	AIG	1/2/2009	1.69	0	
NYSE	AIV	1/2/2009	11.64	0	
NYSE	AIZ	1/2/2009	30.24	0	
NYSE	ALL	1/2/2009	33.26	0	

Query Settings

PROPERTIES

Name: First Trading Day (2)

Description:

APPLIED STEPS

- Source
- ChangedType
- FilteredRows
- RenamedColumns**
- RemovedColumns

LOAD SETTINGS

- ☐ Load to worksheet
- ☐ Load to Data Model

READY PREVIEW DOWNLOADED AT 10:51 AM.

Constraint Search Join

A Constraint Search Join is a join operation which extends a local table with an additional user-specified attribute based on a large corpus of heterogeneous tabular data.

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 €
...

Assumptions on Table Corpus

1. Entity-Attributes-Tables

- One entity per row

2. Subject Attribute

- name of the entity
- string, no number or other data type
- relatively unique values
- used as pseudo-key

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

Input / Output: Constraint Search Join

■ Input:

1. Corpus of heterogeneous tables
2. Query table
3. Subject attribute specification
4. Keywords describing extension attribute

■ Output:

- Query table complemented with extension attribute

Elements of a Search Join

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

- $s()$: Search operator determines the set of relevant tables
- $m()$: MultiJoin operator performs a series of left-outer joins between the query table and all relevant tables
- $c()$: Consolidation operator 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 tables.

■ Dimension of Relevancy

1. Right Attribute

- table should contain the extension attribute
- requires schema matching

2. Entity Coverage

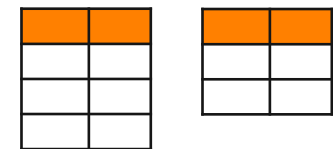
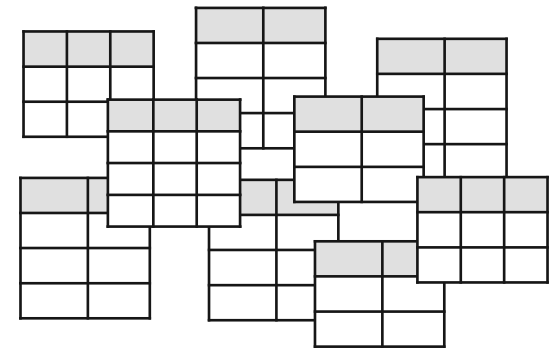
- table should cover many entities from the query table
- requires identity resolution

3. Timeliness

- data should refer to desired point in time

4. Trustworthiness

- data should be trustworthy



Multi-Join Operator

The multi-join operator performs a series of left-outer joins on the subject attribute between the query table and all relevant tables.

■ Input

q = query table

T_r = set of relevant tables

■ Output

t_e = extended query table

No.	Region	GDP per Capita	GDP	GDP (2012)	GDP [current, US\$]
1	Alsace	45.000 €	45.914 €	45.363	48,900 \$
2	Lorraine	NULL	51.233 €	NULL	NULL
3	Guadeloupe	19.000 €	NULL	NULL	20,200 \$
4	Centre	59.500 €	NULL	59.230	63,260 \$

Consolidation Operator

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

■ Input

t_e = extended query table

■ Output

t_r = result table

No	Region	GDP (current, €)
1	Alsace	45.914
2	Lorraine	51.233
3	Guadeloupe	19.000
4	Centre	59.500

■ The operator might employ various conflict resolution functions

- majority vote
- clustered vote
- considering timeliness of data
- considering source trustworthiness
-

Infogather

- Prototype developed by Microsoft Research
- Data corpus: 573 million HTML tables from Bing Crawl (2011)
- Splits HTML Tables into Binary-Entity-Attribute Tables

Subject Attribute

Region	Unemployment
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %

Subject Attribute

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

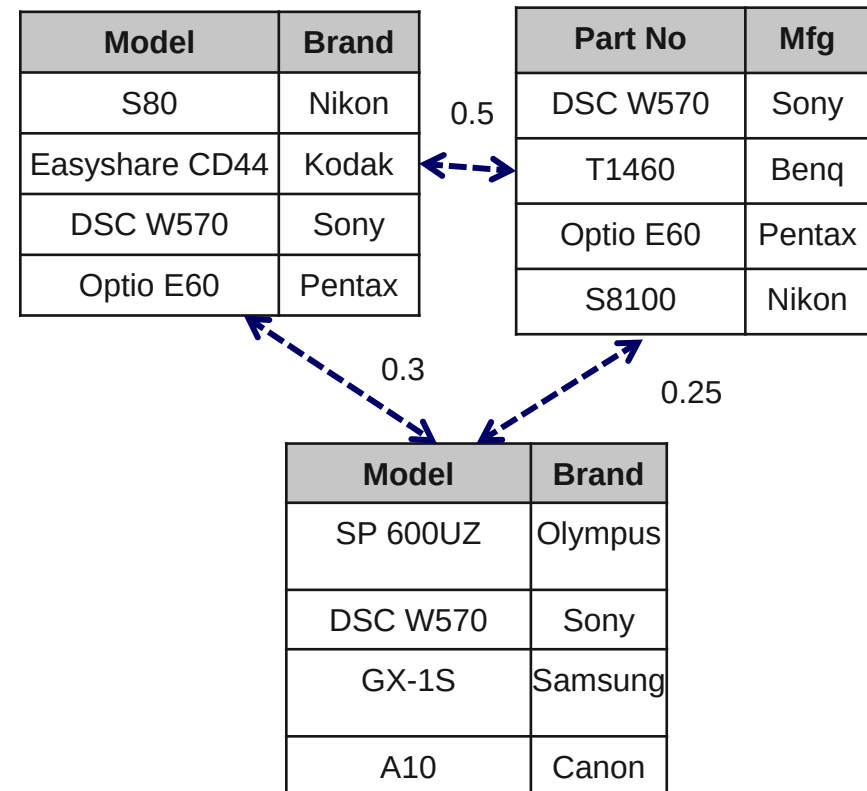
■ Pre-compute matching graph between tables

■ Features used for matching

1. Key values overlap
2. Attribute label similarity
3. Attribute values similarity
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, keywords describing extension attribute
- Web tables considered relevant for the query
 - share subject attribute value with query table
 - and directly match extension attribute
 - **or** are connected to directly matching tables via the matching graph
- Matching score
 - directly matching tables: $\text{entity overlap} / \min(|t_q|, |t_w|)$
 - 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

Query Table

Model	Brand
S80	??
T1460	??
GX-1S	??
A10	??

0.25

Model	Brand
S80	Canon
Easyshare CD44	Kodak
DSC W570	Sony
Optio E60	Pentax

0.4

Part No	Mfg
DSC W570	Sony
T1460	Benq
Optio E60	Pentax
S8100	Nikon

0.25

Model	Brand
SP 600UZ	Olympus
DSC W570	Sony
GX-1S	Samsung
A10	Canon

0.2

Indirect
matching
tables

0.5

Model	Brand
S80	Benq
A10	Innostream
A460	Samsung
S710	HTC

Direct
matching
tables

Evaluation

■ 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

Shortcoming

- Users need to know the extension attribute beforehand
- No discovery of additional relevant attributes

Unconstraint Search Join

Given a local table and a subject 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 department	...
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 Cloud
2. Follows RDF links into other Linked Data sources
3. Provides for mining extended table using all RapidMiner features

Input Table

Link

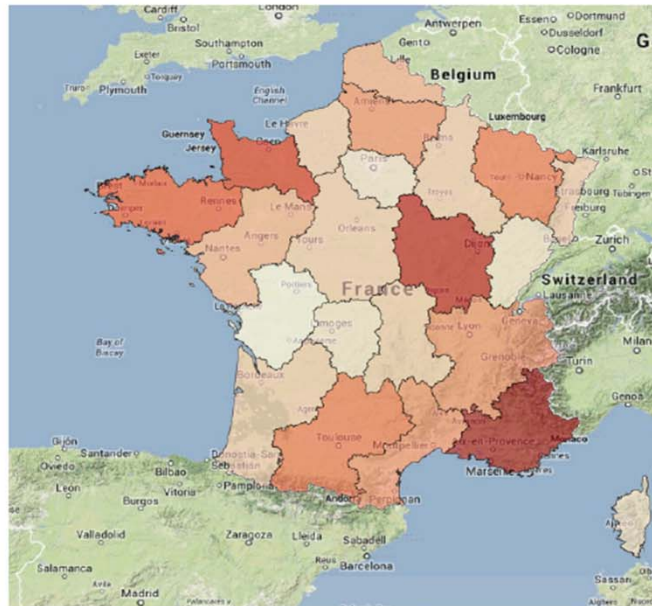
Additional Attributes

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_...	r
18	Bretagne	0.097	http://wifo5-0	71136.200	1097.057	1.220	0.790	15536.500	13490	924	888.400	3704	248	2
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	3
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	4
14	Auvergne	0.102	http://wifo5-0	29729.200	689.348	1.580	0.260	16396.800	6450	1009.400	951.900	2374	151	2
11	Centre	0.107	http://wifo5-0	59502	868.893	1.140	0.340	16880.700	12992	13745.600	761.100	3838	317	3
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	3
21	Midi-Pyréné	0.110	http://wifo5-0	64525	2282.610	2.180	1.160	15728.200	12598	7964	784.500	4820	304	4
3	Franche-Cor	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	8
23	Aquitaine	0.112	http://wifo5-0	74099.700	1146.621	1.070	0.930	16118.900	15200	5894	802.800	6088	349	4
17	Poitou-Char	0.113	http://wifo5-0	37880.700	305.333	0.760	0.620	15766.100	7850	138.500	698.700	2490	239	2
25	Corse	0.114	http://wifo5-0	5686.200	13.209	0.900	0.970	14775.800	858	448	921.100	1115	44	8
2	Lorraine	0.122	http://wifo5-0	51233.500	546.504	0.920	0.160	16035.400	17315	9565.700	824.400	3640	192	2
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	2
24	Provence-Al	0.131	http://wifo5-0	117460.300	2099.396	1.650	0.870	16582.400	27335	5690	813.200	14913	469	3

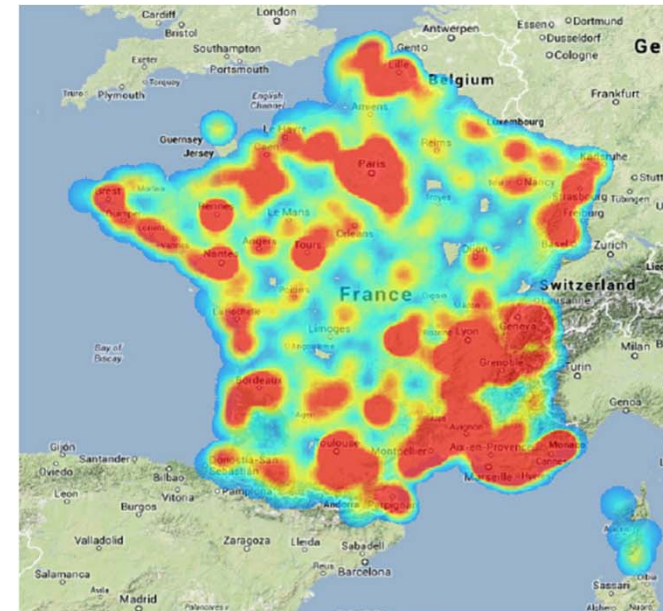
<http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/>

Resulting Correlations

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



(a) Unemployment by region



(b) Heat map of police stations

Petar Ristoski, Christian Bizer, Heiko Paulheim: Mining the Web of Linked Data with RapidMiner. Journal of Web Semantics, 2015.

Shortcoming

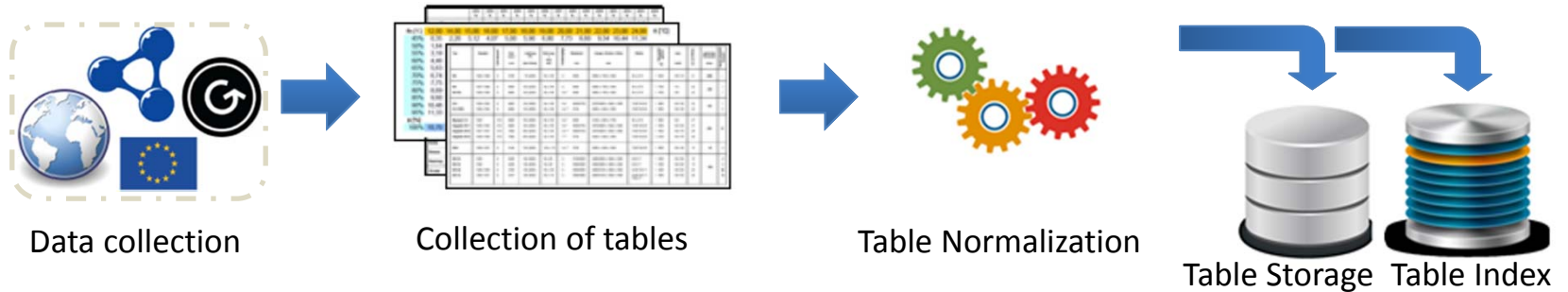
- Requires RDF links
- Not applicable to other forms of Web data
(HTML tables, Microdata annotations, data portals, ...)

3. The Mannheim Search Join Engine (MSJE)

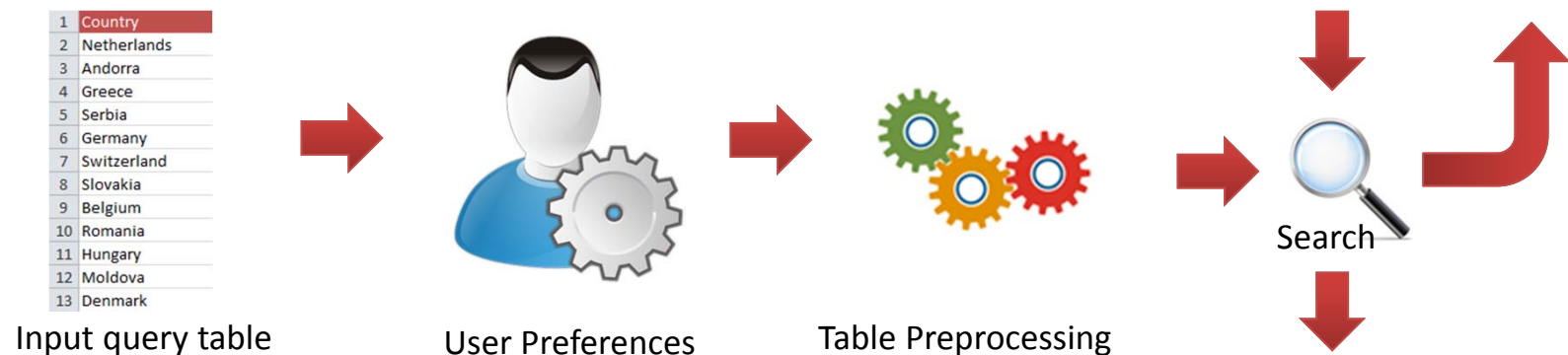
- Provides for constraint and unconstraint search joins

UNIVERSITY OF
MANNHEIM

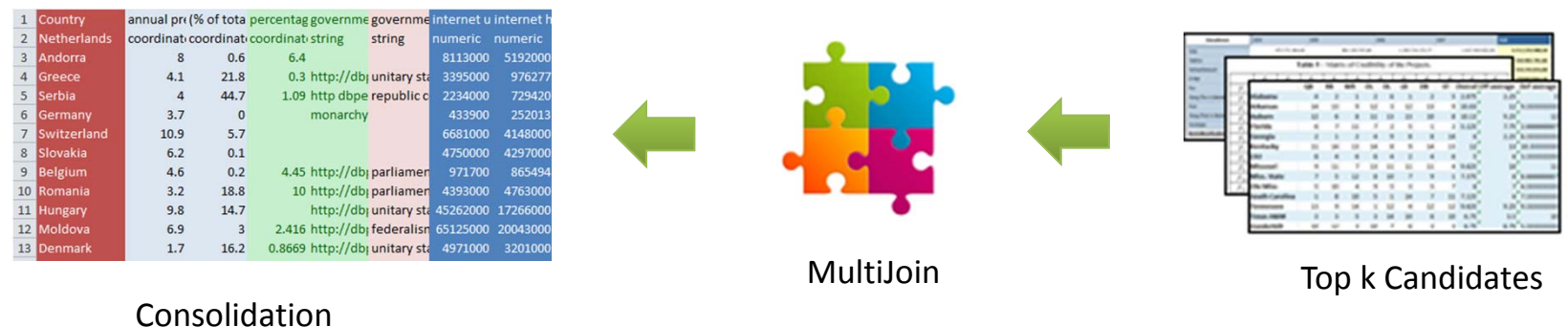
1. Table Indexing



2. Table Search



3. Data Consolidation

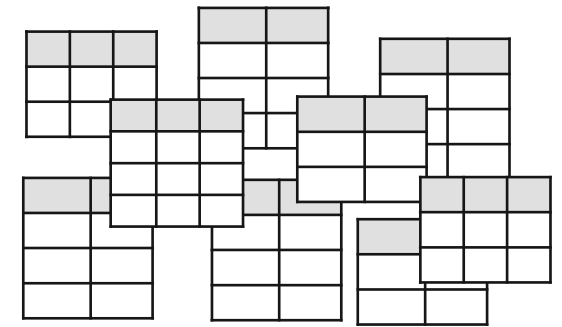


The Search Operator

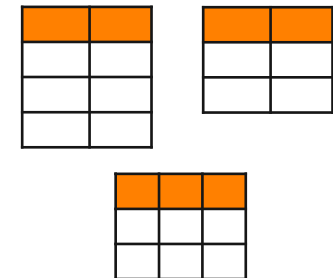
All tables that describe the entities from the query table are relevant.

■ Table Ranking

- Subject column overlap
- Two variations:
 1. Exact match of tokens
 2. Similarity match of tokens using FastJoin library (fuzzy Jaccard)



Relevant



■ Select TopK Tables

- We use Top 1000 tables in the single column experiments

Collective Disambiguation

Query Table

City
Berlin
Madison
Chatham
Perth

Web Table 1

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

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

Web Table 2

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

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

Multi-Join Operator

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

No.	Region	Unemploy	Unemp	GDP	GDP per C
1	Alsace	11 %	NULL	45.914 €	45.000 €
2	Lorraine	12 %	NULL	51.233 €	NULL
3	Guadeloupe	28 %	26 %	NULL	19.000 €
4	Centre	10 %	9.4 %	NULL	59.500 €



Consolidation Operator

The consolidation operator merges corresponding columns and fuses values in order to return a concise result table.

■ Column Matching

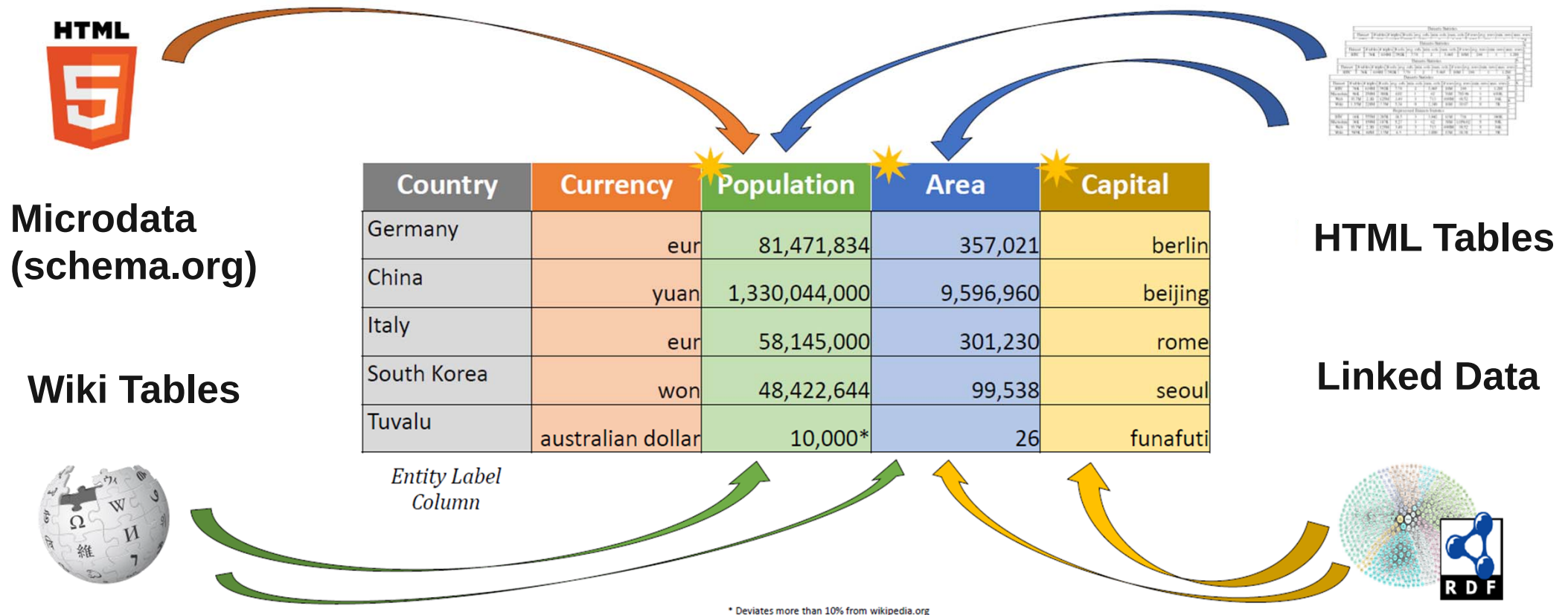
- Combination of label- and instance-based techniques

■ Conflict Resolution

1. cluster similar values
2. choose largest cluster
3. choose most frequent value
(alternative: calculate mean)

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 €

Evaluation: Types of Web Data Used



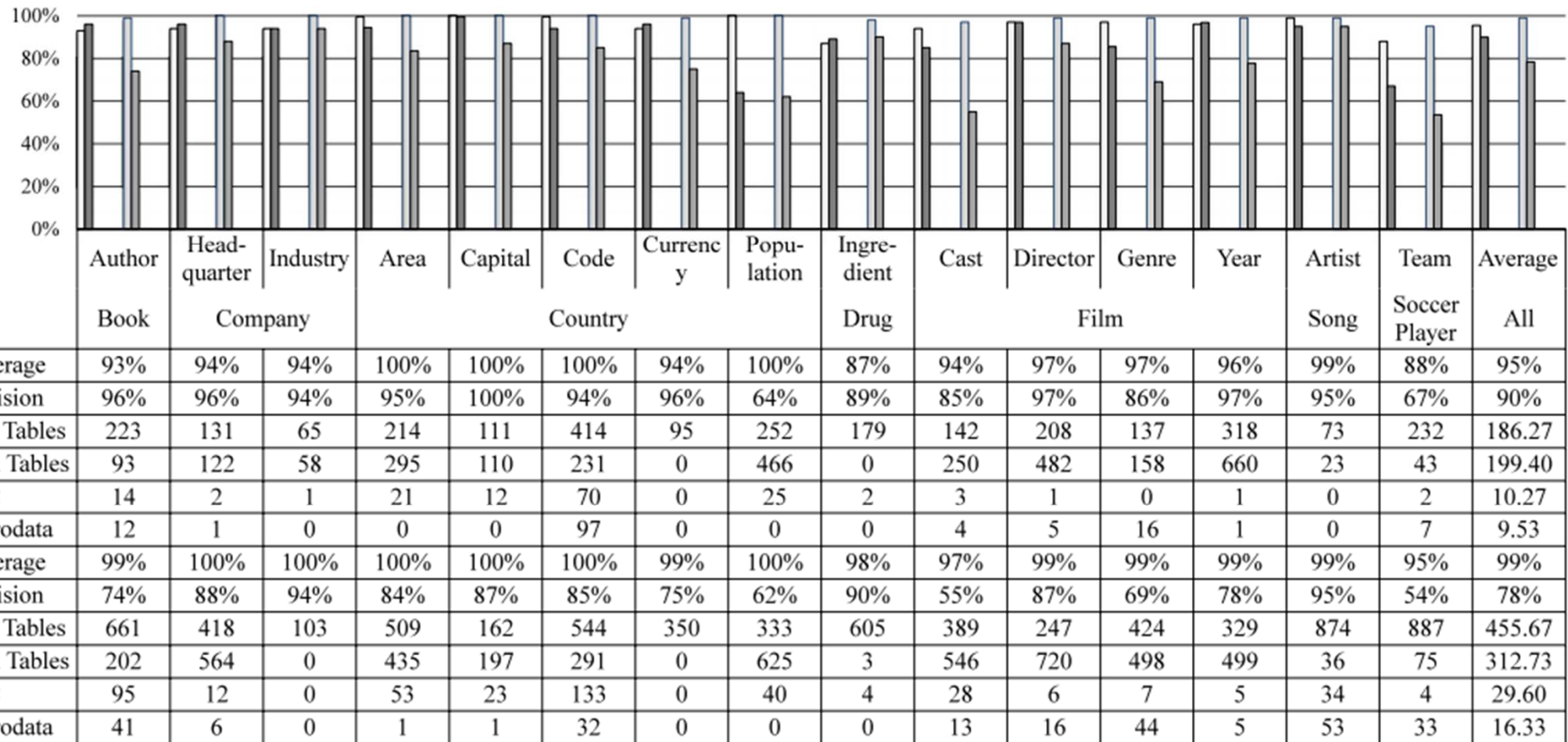
- **Microdata:** Web Data Commons Corpus. <http://webdatacommons.org/structureddata/>
- **HTML Tables:** Web Data Commons Table Corpus. <http://webdatacommons.org/webtables/>
- **Wiki Tables:** Northwestern University Corpus. <http://downey-n1.cs.northwestern.edu/public/>
- **Linked Data:** Billion Triples Challenge 2014: <http://km.aifb.kit.edu/projects/btc-2014/>

Corpus Statistics

Corpora statistics										
Corpus	# Tables	# Triples	# Cols.	Avg. cols.	Min. cols.	Max. cols.	# Rows	Avg. rows	Min. rows	Max. rows
Web tables	147.6M	6.7B	510M	3.49	2	2368	1.8B	12.41	1	70K
Microdata	96K	250M	388K	4.02	1	62	76M	785.96	1	630K
BTC	76K	634M	592K	7.78	2	5465	18M	244.00	1	1.2M
Wiki tables	1.35M	220M	7.5M	5.34	0	2349	16M	10.97	0	5K
Preprocessed corpora statistics										
Web tables	35.7M	2.3B	125M	4.49	3	713	699M	19.52	5	36K
Microdata	36K	150M	187K	5.27	3	62	38M	1056.02	5	50K
BTC	16K	555M	285K	18.5	3	3942	11M	716.00	5	340K
Wiki tables	585K	60M	3.7M	6.3	3	1000	12M	19.38	5	5K

- Data originates from over 1 million different websites
- Tables filtered out in preprocessing
 - less than 5 rows, less than 3 columns
 - no subject attribute

Evaluation Results: Constraint Search Joins



- Size of query tables: 50-200 rows
- Ground truth: IMDB, Wikipedia

Example Result: Extend with Single Column

Song-Artist

Extended Table

Show ▾ entries

Search:

Song	Artist <u>(287 Sources)</u>
string	string
gimme shelter	rolling stones
god only knows	beach boys
good golly miss molly	little richard
good vibrations	beach boys
great balls fire	jerry lee lewis
heartbreak hotel	elvis presley
help	beatles
heroes	david bowie
hey jude	beatles
hotel california	eagles

Showing 21 to 30 of 99 entries

Provenance Summary

Song-Artist

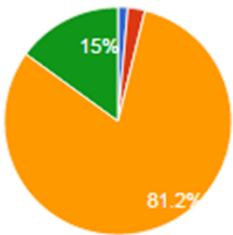
Table Statistics



Total Number of Columns: 2

Total Number of Entries: 100

Table Sources Distribution



- BTC Tables
- Microdata Tables
- Web Tables
- Wikipedia Tables

Search:

Artist (287 Sources)



string



rolling stones

beach boys

little richard

beach boys

jerry lee lewis

elvis presley

beatles

david bowie

beatles

eagles

heartbreak hotel

help

heroes

hey jude

hotel california

Showing 21 to 30 of 99 entries

Show Statistics

Export CSV

Export RDF

Provenance Details

Home Examples Live Demo About Contact Credits

Song-Artist

UNIVERSITY OF
MANNHEIM

Table Sources

BTC sources:

- www.bbc.co.uk_http__purl.org_ontology_po_MusicSegment
- lastfm.rdfize.com_http__purl.org_ontology_mo_Performance
- lastfm.rdfize.com_http__purl.org_ontology_mo_Performance
- www.bbc.co.uk_http__purl.org_ontology_po_MusicSegment

Microdata sources:

- lastfm.se_http__schema.org_MusicRecording
- searchmp3.mobi_http__schema.org_MusicRecording
- qobuz.com_http__schema.org_MusicRecording
- palcomp3.com_http__schema.org_MusicRecording
- songsterr.com_http__schema.org_MusicRecording
- pandora.com_http__schema.org_MusicRecording
- wom.de_http__schema.org_MusicRecording

Wiki sources:

- Full_list_462907_957939
- N_O_329289_636703
- All_Hot_100_singles_349646_689163
- Number-one_singles_529408_1104447

Search:

Artist (287 Sources)

string

rolling stones

beach boys

little richard

beach boys

jerry lee lewis

elvis presley

beatles

david bowie

beatles

eagles

hey jude

hotel california

Showing 21 to 30 of 99 entries

Show Statistics

Export CSV

Export RDF

<http://searchjoins.webdatacommons.org/>

Evaluation Results: Unconstraint Search Joins

Query	<u>Added attributes</u>	<u>Max. fused</u>	Web tables	Wiki tables	Microdata	BTC
Book	2551	37	804	112	47	25
City	2685	70	780	172	13	28
Company	2056	57	920	40	21	16
Country	2328	157	753	233	1	23
Drug	1936	90	923	22	4	48
Film	3194	19	783	150	39	27
Soccer player	1695	72	895	91	14	0
Song	2651	16	888	53	38	19

Parameters

- Minimum attribute density: 50%
- Similarity threshold for merging attributes:
0.8 for string attributes and 0.4 for numeric attributes

Reduced Example: Unconstraint Search Joins

[Home](#) [Examples ▾](#) [Live Demo](#) [About](#) [Contact](#) [Credits](#)

UNIVERSITY OF
MANNHEIM

Countries

Extended Table

Show entries

Search:

country	population (33 Sources)	area (sq. km) (18 Sources)	number of time zones (8 Sources)	lifeexpectancy (10 Sources)	n
string	numeric	numeric	numeric	numeric	n
abkhazia	216000	8600	5	null	4
afghanistan	2.99E+07	647500	1	45.02	0
albania	3170000	28748	1.05	77.41	4
algeria	3.38E+07	2381741	1.01	73.7	4
andorra	78115	468	1	82.43	0
angola	1.59E+07	1246700	1.02	54.14	3
argentina	4.01E+07	2780400	1	76.95	3
australia	20090437	7741220	1.03	81.81	6
bahrain	727785	694	1.33	77.5	3
benin	8532547	112620	1	57.83	3

Showing 1 to 10 of 98 entries

Previous **1** 2 3 4 5 ... 10 Next

Show Statistics

Export CSV

Export RDF

Provenance Summary

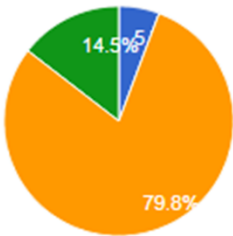
Countries

Table Statistics



Total Number of Columns: 505
Total Number of Entries: 99

Table Sources Distribution



- BTC Tables
- Web Tables
- Wikipedia Tables

	lifeexpectancy (10 Sources)	no of unique maritime boundaries (4 Sources)
	numeric	numeric
angola	54.14	4
argentina	76.95	0
australia	81.81	4
bahrain	77.5	4
benin	57.83	0



Showing 1 to 10 of 98 entries

Summary: 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 WDC Microdata WikiTables data Linked Data

- 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.
- Lehmberg, et al.: The Mannheim Search Join Engine. Journal of Web Semantics 2015.

4. Conclusions and Outlook

■ Conclusions

- Search Joins bring together Web Search and DB Joins
- The prototypes show that simple joins are feasible
- Different types of Web data contribute to different queries

■ Outlook

- Application of search joins in specialized domains
 - knowledge base extension (DBpedia)
 - mining intranet data (together with RapidMiner as industry partner)
- How to deal with the overwhelming number of new attributes?

Idea: Schema Complement

Add attributes while preserving “consistency” of the schema.

- What should be considered “consistent”?
 - Possible 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.
 - Not investigated yet: Combination of “consistency” with other quality dimensions like “density” and “trustworthiness”.
-
- Das Sarma, et al.: Finding related tables. SIGMOD 2012.

Advertisement: Open PhD Position

Project: Querypedia - Query-specific Wikipedia Construction

Skills: We are particularly interested in candidates with a background in one or several of the following areas:

- statistical text processing (e.g., automatic summarization, event extraction and ordering)
- machine learning
- knowledge base construction
- information retrieval
- distributed large-scale processing

Salary: TV-L 13 100%

Contact: Laura Dietz. queripidia-jobs@uni-Mannheim.de