International Semantic Web Conference
Riva del Garda, Italy, 22.10.2014

Semantic Web Challenge – Big Data Track

Extending Tables with Data from over a Million Websites

Oliver Lehmberg, Dominique Ritze, Petar Ristoski,

Kai Eckert, Heiko Paulheim, **Christian Bizer**

UNIVERSITÄT MANNHEIM

Goal

Extend a local table with additional columns using different types of Web data.

Region	Un-employment
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %
Centre	10 %
Martinique	25 %

+

GDP per Capita	Population Growth
45.914 €	0,16 %
51.233 €	-0,05 %
19.810 €	1,34 %
59.502 €	1,76 %
NULL	2,64 %

Operation 1: Extend Local Table with Single Column

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

„GDP per Capita“

Region	Unemployment
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %
Centre	10 %
Martinique	25 %
…	…

+

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

Operation 2: Extend Local Table with Many Columns

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

density >= 0.8

Region	Unemp. Rate
Alsace	11 %
Lorraine	12 %
Guadeloupe	28 %
Centre	10 %
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	…
…	…		…

Types of Web Data Used

Microdata (schema.org)

Wiki Tables

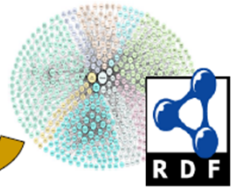
Country	Currency	Population	Area	Capital
Germany	eur	81,471,834	357,021	berlin
China	yuan	1,330,044,000	9,596,960	beijing
Italy	eur	58,145,000	301,230	rome
South Korea	won	48,422,644	99,538	seoul
Tuvalu	australian dollar	10,000*	26	funafuti

Entity Label Column

* Deviates more than 10% from wikipedia.org

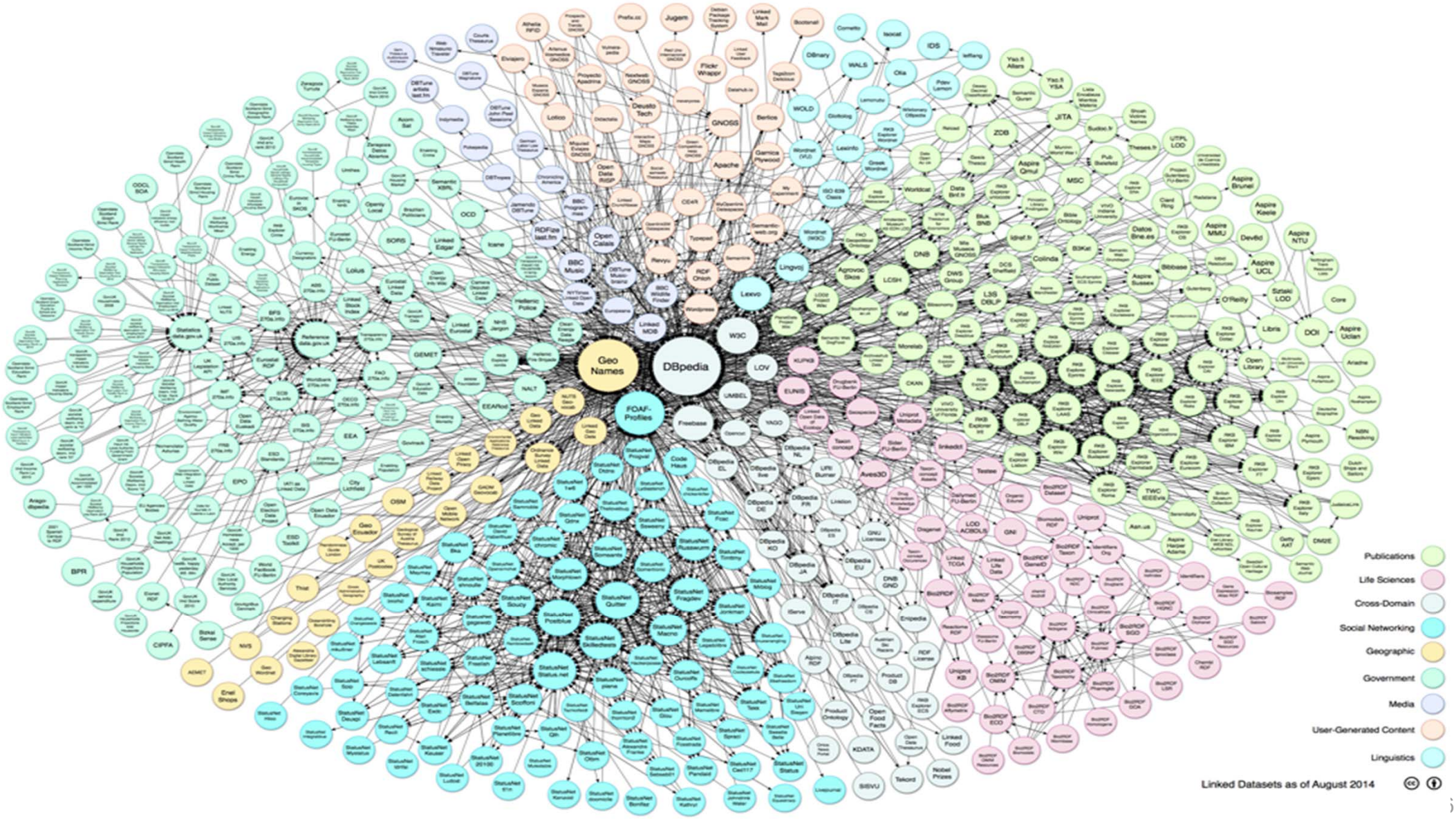
HTML Tables

Linked Data



Billion Triple Challenge Dataset 2014

4 billion triples crawled from **47,000** websites.



Web Data Commons - Microdata Corpus

250 million triples from **463,000** websites.

- Extracted from Common Crawl 2013 web corpus
  - 2.2 billion HTML pages from 12.8 million websites
- Mostly using the schema.org vocabulary
- Main topics
  - Products
  - Reviews
  - Organisations / LocalBusiness
  - Events

Download: http://webdatacommons.org/structureddata/

Web Data Commons – Web Tables Corpus

Around 1% of all HTML tables contain structured data.

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	
Berlin	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
		Northrhine-Westfalia	965,300
		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

- we used **35 million** English HTML tables.
  - extracted from the Common Crawl 2012 web corpus
  - selected out of 11.2 billion raw tables

Web Data Commons – Web Tables Corpus

Column Statistics

Column	#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

Subject Column 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

Download: http://webdatacommons.org/webtables/

WikiTables

1.4 million tables from English Wikipedia.

- extracted by Northwestern University
- from the 2013 Wikipedia XML dump
- only tables, no infoboxes

Download: http://downey-n1.cs.northwestern.edu/public/

Internal Data Model: Entity-Attributes-Tables

- One entity per row
- Subject Column = Name of the entity
  - HTML tables: Most unique string column, break ties by taking leftmost.

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

- Table generation from Linked Data and Microdata
  - generate one table per class and website
  - subject column: rdfs:label, foaf:name, x:name
  - we exploit common vocabularies

Indexed Tables

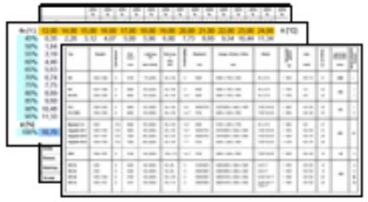
- Selection Conditions:
  1. Minimum size of 3 columns and 5 rows
  2. Subject column detection successful
- Total # of tables: 36.3 million
- Total # of PLDs: ~ 1.5 million
- Total # of triples: 3.0 billion

Datasets Statistics										
Dataset	# tables	# triples	# cols.	avg. cols.	min. cols.	max. cols.	# rows	avg. rows	min. rows	max. rows
BTC	16K	555M	285K	18.5	3	3,942	11M	716	5	340K
Microdata	36K	150M	187K	5.27	3	62	38M	1,056.02	5	50K
Web	35.7M	2.3B	125M	3.49	3	713	699M	19.52	5	36K
Wiki	585K	60M	3.7M	6.3	3	1,000	12M	19.38	5	5K

The Mannheim Search Joins Engine (MSJE)

1. Table Indexing

Data collection



Collection of tables

Table Normalization

Table Storage Table Index

2. Table Search

	Country
1	Netherlands
2	Andorra
3	Greece
4	Serbia
5	Germany
6	Switzerland
7	Slovakia
8	Belgium
9	Romania
10	Hungary
11	Moldova
12	Denmark

Input query table

User Preferences

Table Preprocessing

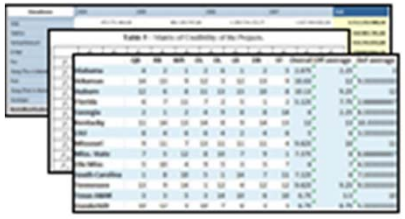
Search

3. Data Consolidation

	Country	annual pr	(% of tota	percentag	governme	governme	internet u	internet h
	Netherlands	coordinat	coordinat	coordinat	string	string	numeric	numeric
3	Andorra	8	0.6	6.4			8113000	5192000
4	Greece	4.1	21.8	0.3	http://db	unitary st	3395000	976277
5	Serbia	4	44.7	1.09	http dbpe	republic c	2234000	729420
6	Germany	3.7	0		monarchy		433900	252013
7	Switzerland	10.9	5.7				6681000	4148000
8	Slovakia	6.2	0.1				4750000	4297000
9	Belgium	4.6	0.2	4.45	http://db	parliamen	971700	865494
10	Romania	3.2	18.8	10	http://db	parliamen	4393000	4763000
11	Hungary	9.8	14.7		http://db	unitary st	45262000	17266000
12	Moldova	6.9	3	2.416	http://db	federalisn	65125000	20043000
13	Denmark	1.7	16.2	0.8669	http://db	unitary st	4971000	3201000

Consolidation

MultiJoin

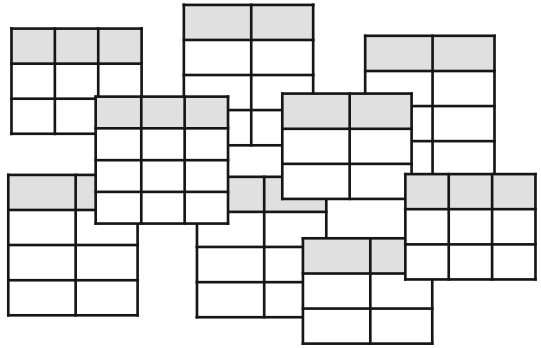


Top k Candidates

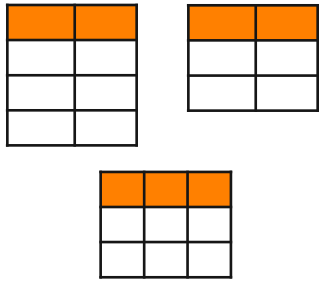
The Search Operator

The Search operator determines the set of relevant Web tables.

- Table Ranking
  - subject column value overlap
  - extended Jaccard Similarity (FastJoin)
- Select TopK Tables
  - 1000 tables in the single column experiments



Relevant



Multi-Join Operator

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

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 €

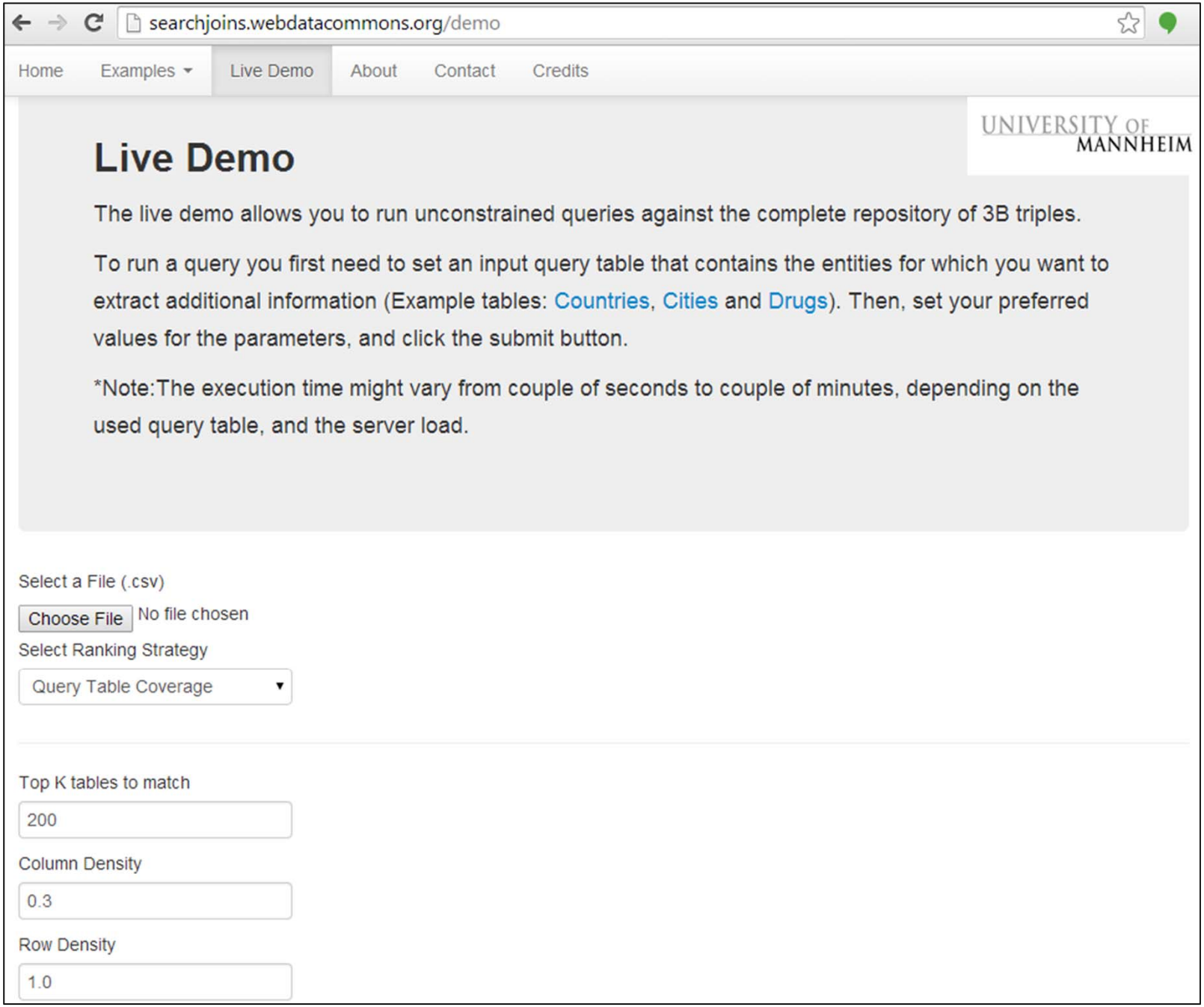
Consolidation Operator

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

- Column Matching
  - Combination of label- and instance-based techniques
- Conflict Resolution
  - Strings: majority vote
  - Numeric values: average, median, clustering and vote

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

http://searchjoins.webdatacommons.org



Result: Extend with Single Column

Home Examples Live Demo About Contact Credits

UNIVERSITY OF MANNHEIM

Song-Artist

Extended Table

Show 10 entries

Search:

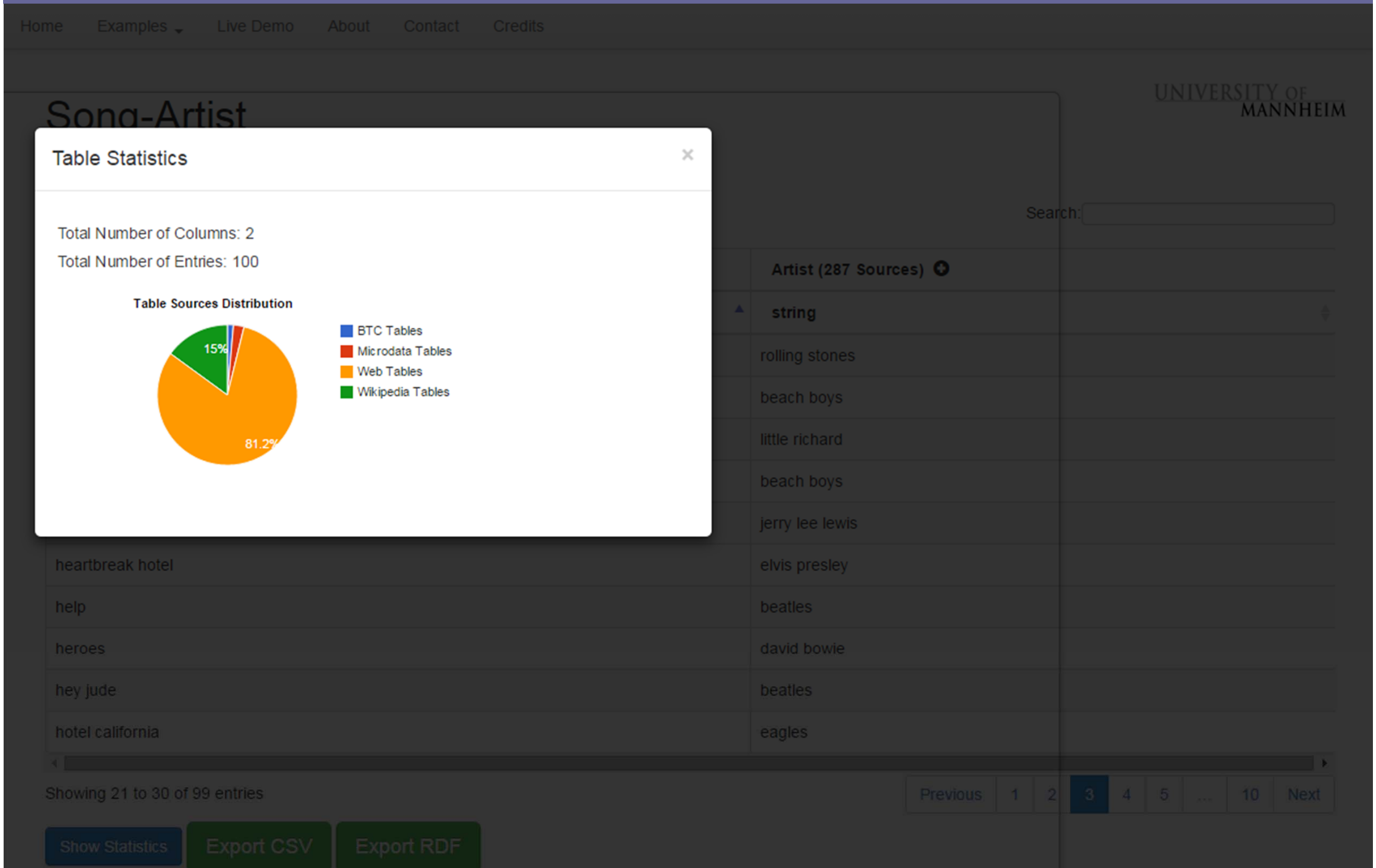
Song	Artist (287 Sources)
string	string
gimme shelter	rolling stones
god only knows	beach boys
good golly miss molly	little richard
good vibrations	beach boys
great balls fire	jerry lee lewis
heartbreak hotel	elvis presley
help	beatles
heroes	david bowie
hey jude	beatles
hotel california	eagles

Showing 21 to 30 of 99 entries

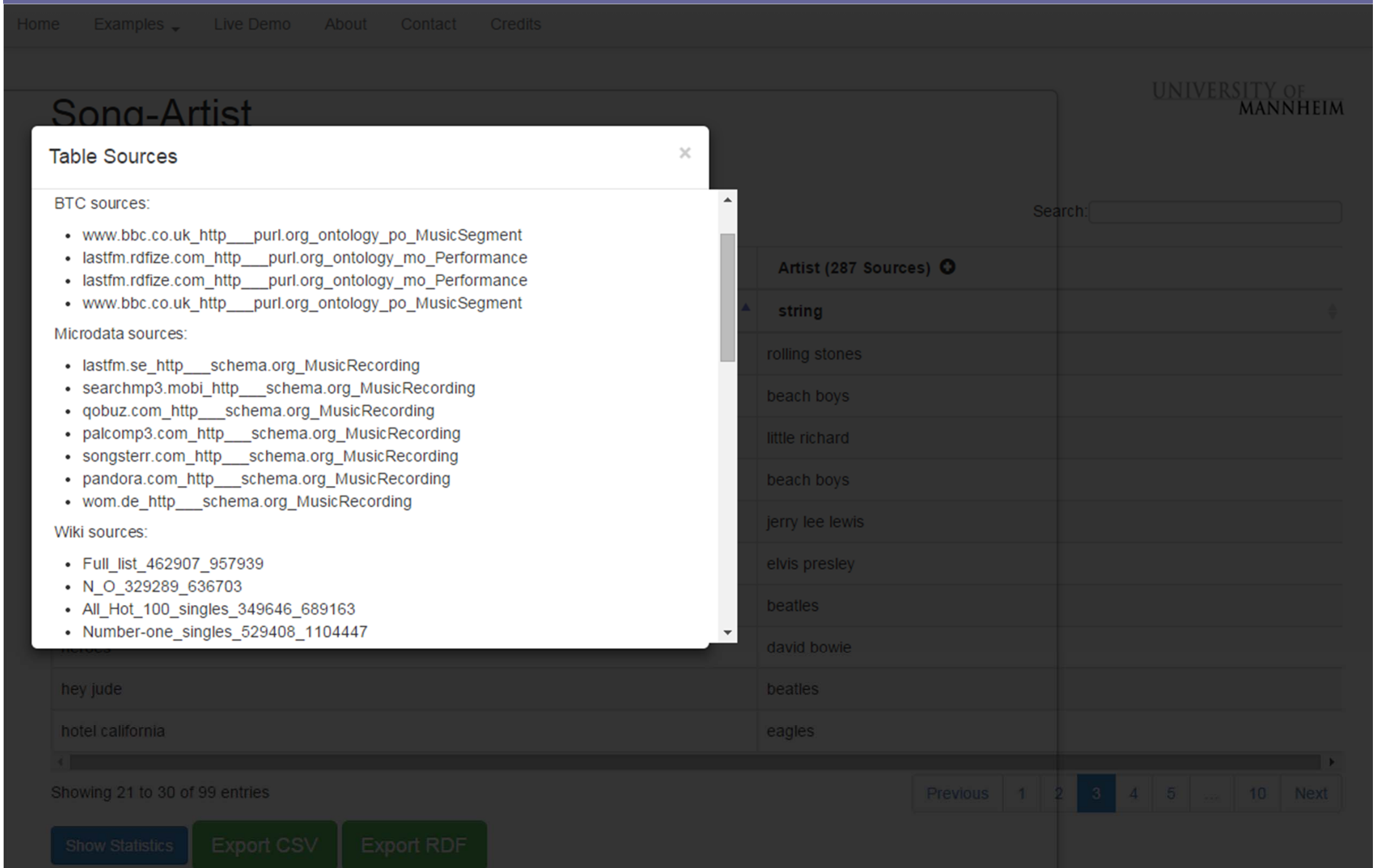
Previous 1 2 3 4 5 … 10 Next

Show Statistics Export CSV Export RDF

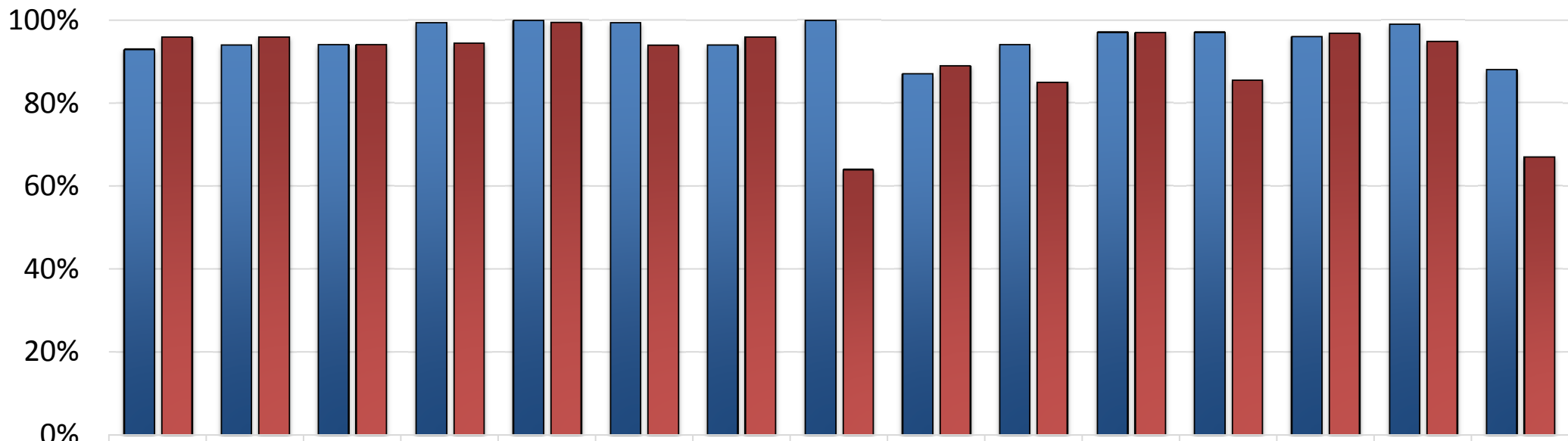
Provenance Summary



Provenance Details



Evaluation Results



	Author	Head-quarter	Industry	Area	Capital	Code	Currency	Popu-lation	Ingre-dient	Cast	Director	Genre	Year	Artist	Team
	Book	Company		Country					Drug	Film				Song	Soccer Player
coverage	93%	94%	94%	100%	100%	100%	94%	100%	87%	94%	97%	97%	96%	99%	88%
precision	96%	96%	94%	95%	100%	94%	96%	64%	89%	85%	97%	86%	97%	95%	67%

Coverage: Percentage of entities for which a value was found.
Precision: Manually evaluated using Wikipedia, IMDB, Amazon.

Result: Extend with Many Columns

Home Examples Live Demo About Contact Credits

UNIVERSITY OF MANNHEIM

Countries

505 columns are added
and filled with data from 2071 tables.

Extended Table

Show 10 entries Search:

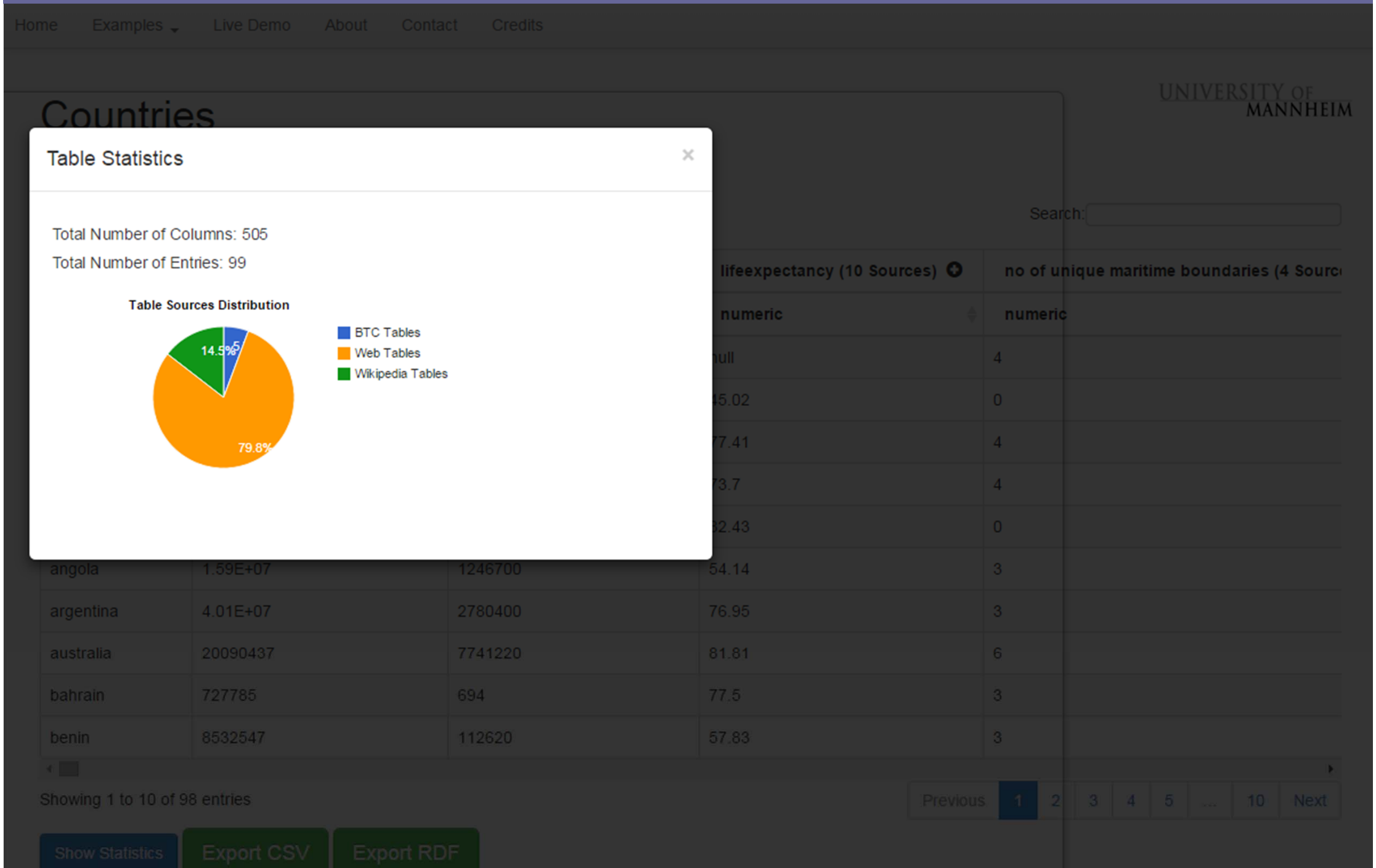
country	population (33 Sources)	area (sq. km) (18 Sources)	number of time zones (8 Sources)	lifeexpectancy (10 Sources)	n
string	numeric	numeric	numeric	numeric	n
abkhazia	216000	8600	5	null	4
afghanistan	2.99E+07	647500	1	45.02	0
albania	3170000	28748	1.05	77.41	4
algeria	3.38E+07	2381741	1.01	73.7	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

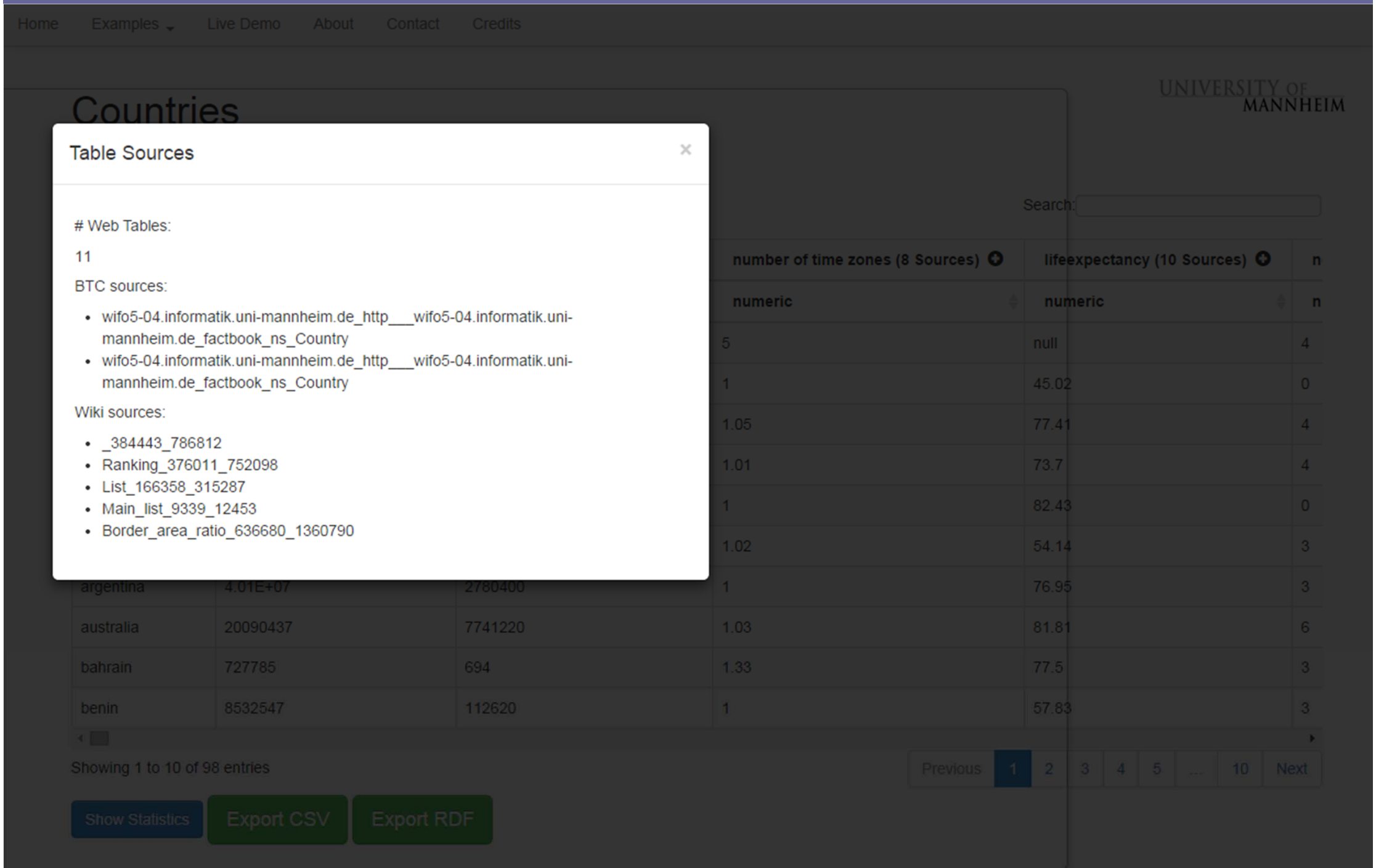
Previous 1 2 3 4 5 … 10 Next

Show Statistics Export CSV Export RDF

Provenance Summary



Provenance Details for “area (sq. km)”



Conclusion

Search Joins bring together Web Search and DB Joins.

- The prototype shows that simple queries are feasible.
- The Web is one application domain for search joins, corporate intranets are the other.
- The overlooked Big Data Vs: Variety and Veracity