UNIVERSITÄT MANNHEIM

Lernen, Wissen, Daten, Analysen (LWDA)
Hasso Plattner Institute, Potsdam
13.9.2016

# Exploring the Application Potential of Relational Web Tables

**Prof. Dr. Christian Bizer**

# Hello

**Professor Christian Bizer**

University of Mannheim

Research Topics

- Web Technologies
- Web Data Integration
- Web Data Profiling

# Data and Web Science Group @ University of Mannheim

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





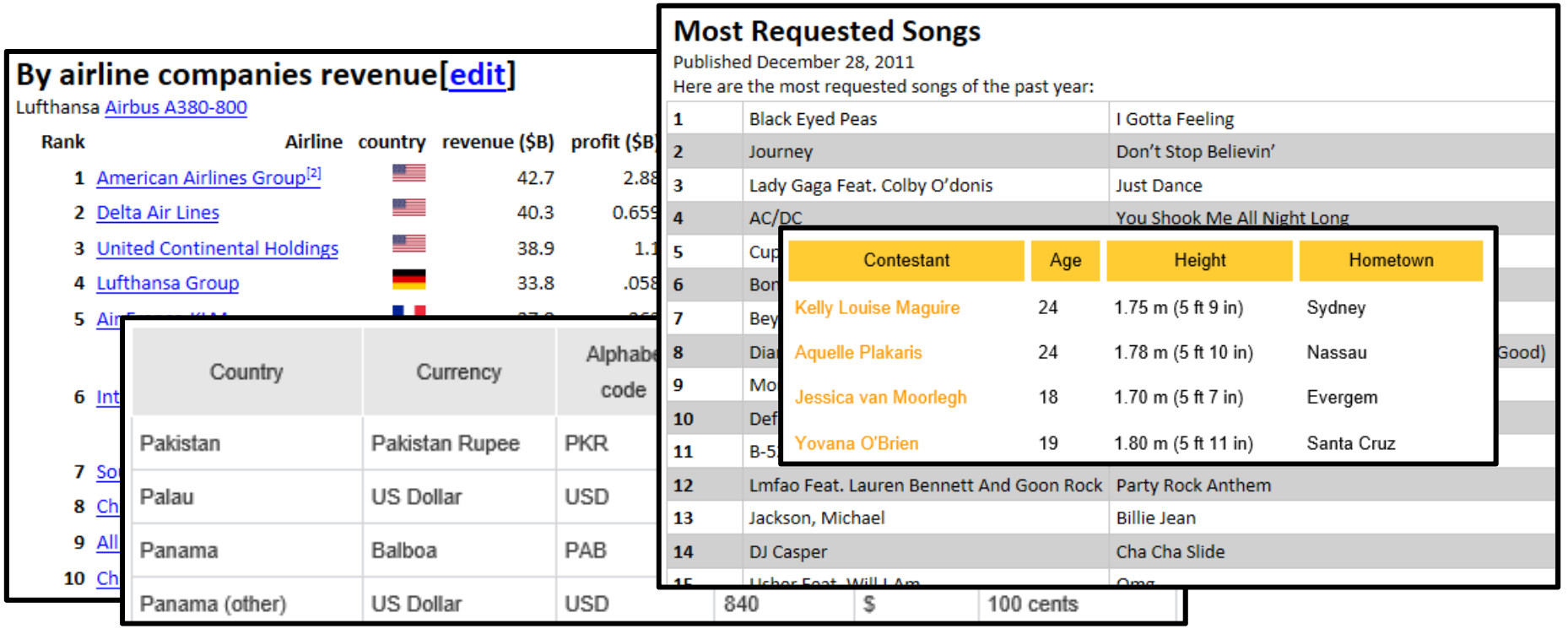
- http://dws.informatik.uni-mannheim.de/

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

# Application Potential of Relational Web Tables

## Main applications so far

1. **Table Augmentation**
2. **Data Translation**



**By airline companies revenue[edit]**

Lufthansa Airbus A380-800

| Rank | Airline | country | revenue ($B) | profit ($B) |
|---|---|---|---|---|
| 1 | American Airlines Group[2] | | 42.7 | 2.88 |
| 2 | Delta Air Lines | | 40.3 | 0.659 |
| 3 | United Continental Holdings | | 38.9 | 1.1 |
| 4 | Lufthansa Group | | 33.8 | .058 |
| 5 | Air | | | |
| 6 | Int | | | |
| 7 | So | | | |
| 8 | Ch | | | |
| 9 | All | | | |
| 10 | Ch | | | |

| Country | Currency | Alphabe code |
|---|---|---|
| Pakistan | Pakistan Rupee | PKR |
| Palau | US Dollar | USD |
| Panama | Balboa | PAB |
| Panama (other) | US Dollar | USD |

**Most Requested Songs**

Published December 28, 2011

Here are the most requested songs of the past year:

| | | |
|---|---|---|
| 1 | Black Eyed Peas | I Gotta Feeling |
| 2 | Journey | Don't Stop Believin' |
| 3 | Lady Gaga Feat. Colby O'donis | Just Dance |
| 4 | AC/DC | You Shook Me All Night Long |
| 5 | Cup | |
| 6 | Bon | |
| 7 | Bey | |
| 8 | Dia | Good) |
| 9 | Mo | |
| 10 | Def | |
| 11 | B-5 | |
| 12 | Lmfao Feat. Lauren Bennett And Goon Rock | Party Rock Anthem |
| 13 | Jackson, Michael | Billie Jean |
| 14 | DJ Casper | Cha Cha Slide |
| 15 | Usher Feat. Will I Am | Omg |

| Contestant | Age | Height | Hometown |
|---|---|---|---|
| Kelly Louise Maguire | 24 | 1.75 m (5 ft 9 in) | Sydney |
| Aquelle Plakaris | 24 | 1.78 m (5 ft 10 in) | Nassau |
| Jessica van Moorlegh | 18 | 1.70 m (5 ft 7 in) | Evergem |
| Yovana O'Brien | 19 | 1.80 m (5 ft 11 in) | Santa Cruz |

| | | | |
|---|---|---|---|
| 840 | $ | 100 cents | |

# Table Augmentation Type 1: Create New Attributes

**Goal: Extend given table with additional attributes and fill attributes with values from the web tables.**

**„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 € |
| … |

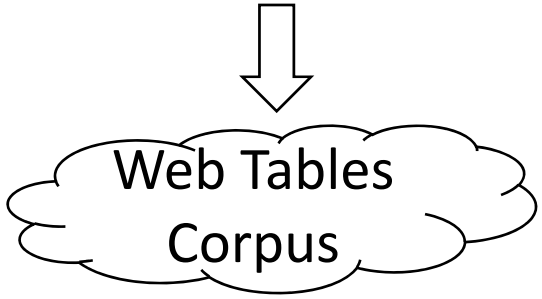
- Cafarella, Halevy, et al.: WebTables: Exploring the Power of Tables on the Web. VLDB 2008.
- Yakout, et al.: InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables. SIGMOD 2012.
- Lehmberg, et al.: The Mannheim Search Join Engine. Journal of Web Semantics 2015.

# Table Augmentation Type 2: Fill Missing Values

- Interesting for cross-domain knowledge bases
- Easier as more existing knowledge can be exploited



| Country | Capital | Population |
|---|---|---|
| Germany | Berlin | |
| France | | 64,000,000 |
| United Kingdom | London | 60,900,000 |
| Canada | | |
| USA | Washington D.C. | |
| Mexico | Mexico City | 109,900,00 |




| Country | Capital | Population |
|---|---|---|
| Germany | Berlin | 82,000,000 |
| France | Paris | 64,000,000 |
| United Kingdom | London | 61,000,000 |
| Canada | Ottawa | 33,000,000 |
| USA | Washington D.C. | 304,000,000 |
| Mexico | Mexico City | 110,000,00 |

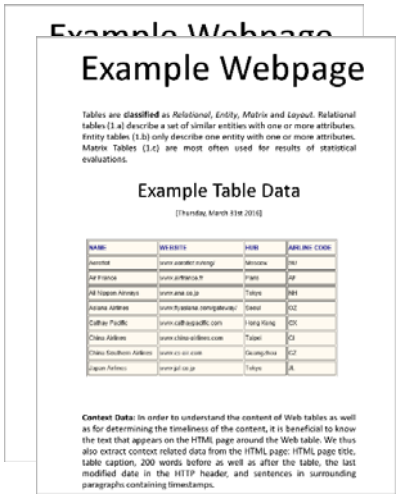
- Dong, et al.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. KDD 2014.
- Ritze, et al: Profiling the Potential of Web Tables for Augmenting Knowledge Bases. WWW 2016.

# The Table Augmentation Process

## 1. Extraction

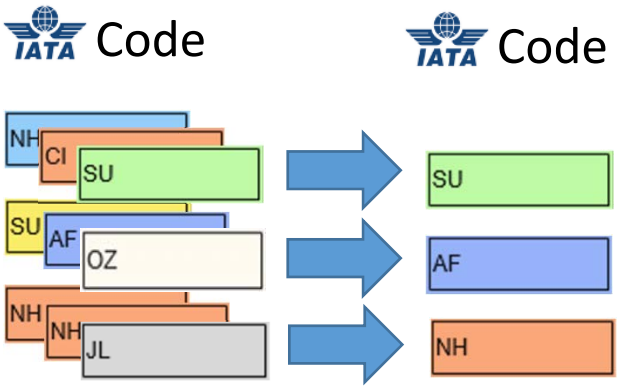
## 2. Matching

## 3. Fusion









# Outline

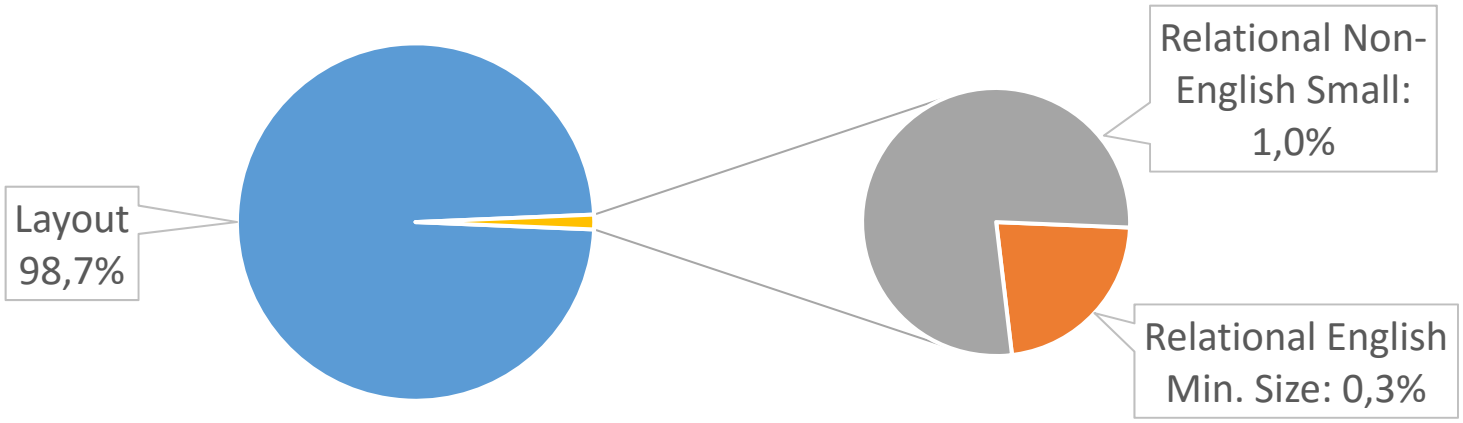
1. WDC Web Table Corpus
2. Matching the WDC Corpus to DBpedia
3. Fusing Web Table Data
4. Lessons Learned

# 1. Web Data Commons (WDC) Web Tables Corpus

- Early research used tables from Google and Bing crawls
  - Cafarella/Halevy (2008): In corpus of 14B raw tables, 154M are "good" relations (1.1%).
  - Yakout, et al. (2012): 650M single-attribute tables
- Problem: Crawls/tables not public, research not verifiable
- Common Crawl enabled public research in this area
  - Series of 1.8-3.5 billion page public web crawls, since 2012
- Public Web Table Corpora
  - WDC Web Tables Corpus 2012: 147 million web tables
  - Dresden Web Tables Corpus 2014: 125 million web tables
  - WDC Web Tables Corpus 2015: 233 million web tables
- http://webdatacommons.org/webtables

# Table Extraction & Classification

1. Common Crawl 2012
   - 3,3b HTML pages
   - from 40m PLDs
2. HTML Table Extraction
   - 11b HTML tables
3. Layout vs. Relational Table Classification
   - 147m relational tables (1.3%)
4. Filtering by Size & Language
   - At least three columns & five rows
   - Only English language
   - 33m resulting tables



# Most Frequent Websites

| Website | # Tables | Topic |
|---|---|---|
| apple.com | 50,910 | Music |
| baseball-reference.com | 25,647 | Sports |
| latestf1news.com | 17,726 | Sports |
| nascar.com | 17,465 | Sports |
| amazon.com | 16,551 | Products |
| wikipedia.org | 13,993 | Various |
| inkjetsuperstore.com | 12,282 | Products |
| flightmemory.com | 8,044 | Flights |
| windshieldguy.com | 7,305 | Products |
| citytowninfo.com | 6,293 | Cities |
| blogspot.com | 4,762 | Various |
| 7digital.com | 4,462 | Music |

# Types of Web Tables

## 1. Relational Tables

| | Lake | Area |
|---|---|---|
| 1 | Windermere | 5.69 sq mi (14.7 km$^2$) |
| 2 | Kielder Reservoir | 3.86 sq mi (10.0 km$^2$) |
| 3 | Ullswater | 3.44 sq mi (8.9 km$^2$) |
| 4 | Bassenthwaite Lake | 2.06 sq mi (5.3 km$^2$) |
| 5 | Derwent Water | 2.06 sq mi (5.3 km$^2$) |

## 2. Entity Tables

| | |
|---|---|
| **Government**[3] | |
| • **Type** | Mayor–Council |
| • **Body** | New York City Council |
| • **Mayor** | Bill de Blasio (D) |
| **Area**[2] | |
| • **Total** | 468.9 sq mi (1,214 km$^2$) |
| • **Land** | 304.8 sq mi (789 km$^2$) |

## 3. Matrix

| | Right-handed | Left-handed | Total |
|---|---|---|---|
| **Males** | 43 | 9 | 52 |
| **Females** | 44 | 4 | 48 |
| **Totals** | 87 | 13 | 100 |

| Table Types in WDC 2015 Corpus | | |
|---|---|---|
| **#Type** | **#Tables** | **% of all tables** |
| **Relational** | 90,266,223 | 0.90 |
| **Entity** | 139,687,207 | 1.40 |
| **Matrix** | 3,086,430 | 0.03 |
| **Sum** | **233,039,860** | **2.25** |

- Eberius, et al.: Building the Dresden Web Table Corpus: A Classification Approach. BDC 2015.
- Qiu, et al.:, DEXTER: Large-Scale Discovery and Extraction of Product Specifications, VLDB 2015.

# Assumptions of the Existing Extension Algorithms

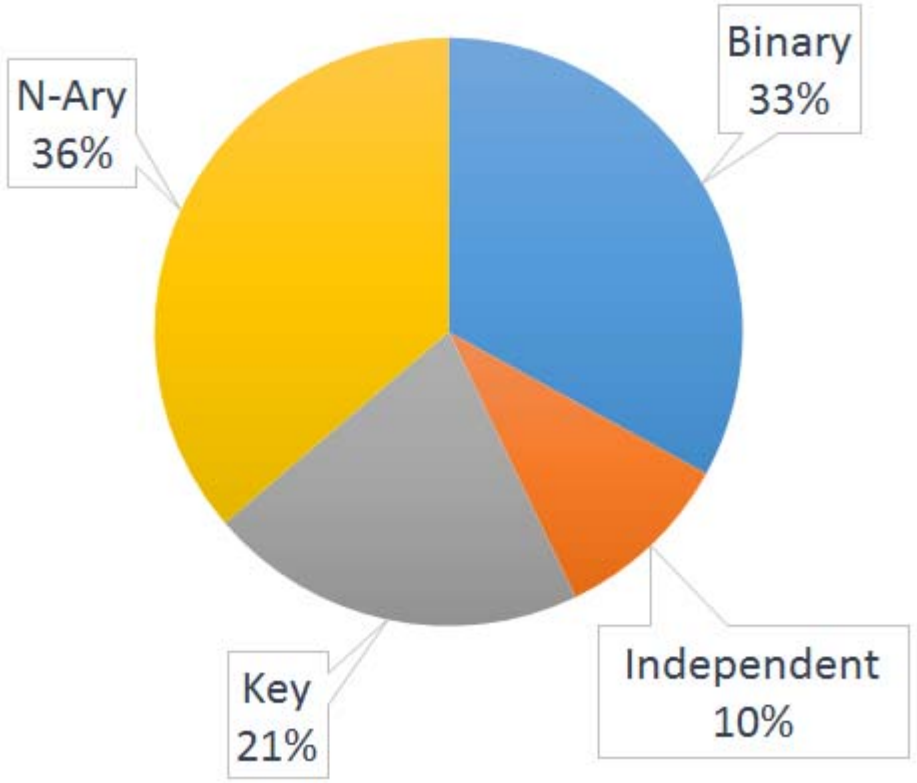
1. Input is corpus of relational tables
   - One entity per row
2. Each table has a subject column
   - name of the entity
   - string, no number or other data type
   - used as pseudo-key
   - accuracy of automatic subject column detection: >90%

| 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 |

# Attribute Dependency on Subject Column

- Manual annotation of 400 relational tables (1,814 columns)

- Binary
  - Attribute depends only on subject column (key)
- N-Ary
  - Attribute depends on subject key and other partial keys contained on the page around the table
  - e.g. type or date of competition in sports results




Lehmberg, et al.: Web Table Column Categorisation and Profiling. WebDB 2016.

# 2. Table Matching

- T2K Matching Framework creates
  - Table-to-Class correspondences
  - Row-to-Instance correspondences
  - Column-to-Property correspondences
- Size of DBpedia (2014)
  - 680 classes
  - 2700 properties
  - 4.5 million instances

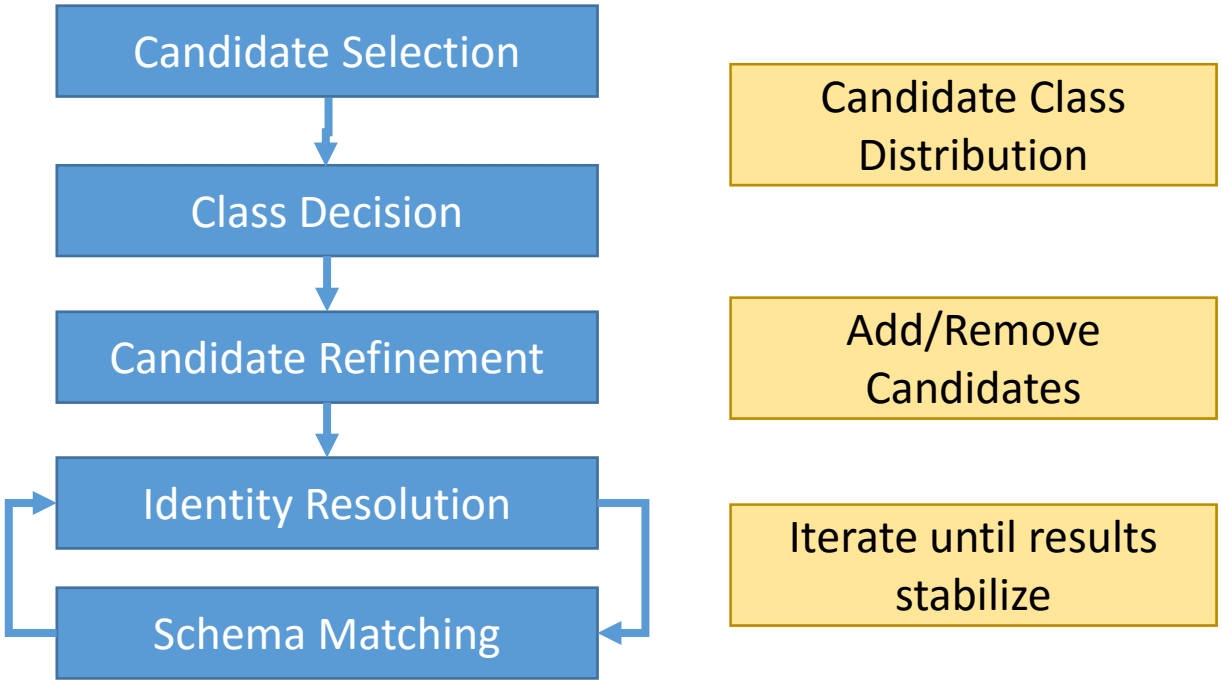


DBpedia:VideoGame

| | Year | Game | DBpedia:Developer Company |
|---|---|---|---|
| DBpedia:Portal | 2007 | *Portal* | Valve Corporation |
| | 2008 | *Fallout 3* | Bethesda Game Studios |
| | 2009 | *Uncharted 2: Among Thieves* | Naughty Dog |
| | 2010 | *Red Dead Redemption* | Rockstar San Diego |
| | 2011 | *The Elder Scrolls V: Skyrim* | Bethesda Game Studios |
| | 2012 | *Journey* | Thatgamecompany |
| | 2013 | *The Last of Us* | Naughty Dog |
| | 2014 | *Middle-earth: Shadow of Mordor* | Monolith Productions |
| | 2015 | *The Witcher 3: Wild Hunt* | CD Projekt RED |

Ritze, et al.: Matching HTML Tables to DBpedia. WIMS 2015.

# T2K Table Matching Algorithm




Tested on gold standard of 233 tables

- 26,124 instance correspondences
- 653 property correspondences

| Task | Precision | Recall | F1 |
|---|---|---|---|
| Instance | .90 | .76 | .82 |
| Property | .77 | .65 | .70 |
| Class | .94 | .94 | .94 |

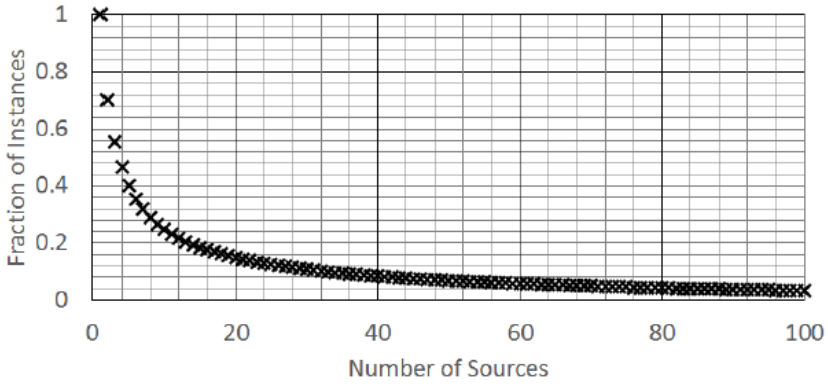
Ritze, et al.: Matching HTML Tables to DBpedia. WIMS 2015.

# Table Matching Results

- Approx. 1 million tables match DBpedia (~3%)
  - 13,726,582 instance correspondences
  - 562,445 property correspondences
  - 301,450 tables with property correspondences (ca. 32%)
  - = 8 million triples

- Content variety
  - 274 different classes (40% of DBpedia)
  - 721 unique properties (26% of DBpedia)
  - 717,174 unique instances (15.6% of DBpedia)

- Head vs. tail instances
  - 30% appear only once
  - 25% appear at least in 10 sources
  - 3% appear in more than 100 sources



# Table Matching – Detailed Results

| DBpedia Class | Number of Tables/Values | | | Number of Values per Data Type | | | |
|---|---|---|---|---|---|---|---|
| | Tables | T. w/ property | Values | Numeric | Date | String | Reference |
| Person | 265 685 | 103 801 | 4 176 370 | 2 117 793 | 1 588 475 | 266 628 | 203 474 |
| Athlete | **243 322** | 95 916 | **3 861 641** | 2 084 017 | 1 435 775 | 163 771 | 178 078 |
| Artist | 9 981 | 2 356 | 18 886 | 3 | 11 527 | 3 499 | 3 857 |
| Politician | 3 701 | 1 388 | 18 505 | 10 | 7 725 | 3 393 | 7 377 |
| Office Holder | 2 178 | 1 435 | 131 633 | 30 | 66 762 | 59 332 | 5 509 |
| Organisation | **194 317** | 36 402 | **573 633** | 99 714 | 187 370 | 100 710 | 185 839 |
| Company | 97 891 | 6 943 | 203 899 | 58 621 | 83 001 | 34 665 | 27 612 |
| Sports Team | 50 043 | 2 722 | 31 866 | 2 206 | 22 368 | 43 | 7 249 |
| Educational Inst. | 25 737 | 14 415 | 238 365 | 38 056 | 64 578 | 13 334 | 122 397 |
| Broadcaster | 14 515 | 11 315 | 93 042 | 564 | 13 095 | 52 186 | 27 197 |
| Work | **269 570** | 127 677 | **2 284 916** | 109 265 | 1 354 923 | 33 091 | 787 637 |
| Musical Work | 138 676 | 80 880 | 1 131 167 | 64 545 | 396 940 | 7 610 | 662 072 |
| Film | 43 163 | 9 725 | 256 425 | 10 844 | 198 913 | 14 382 | 32 286 |
| Software | 39 382 | 23 829 | 486 868 | 418 | 414 092 | 9 194 | 63 164 |
| Place | **133 141** | 24 341 | **859 995** | 413 375 | 273 510 | 84 111 | 88 999 |
| Populated Place | 119 361 | 21 486 | 787 854 | 405 406 | 257 780 | 57 064 | 67 604 |
| Country | 36 009 | 6 556 | 208 886 | 93 107 | 66 492 | 31 793 | 17 494 |
| Settlement | 17 388 | 2 672 | 17 585 | 4 492 | 6 662 | 2 444 | 3 987 |
| Region | 12 109 | 427 | 5 625 | 3 097 | 897 | 292 | 1 339 |
| Architect. Struct. | 10 136 | 1 815 | 46 067 | 3 976 | 7 387 | 23 110 | 11 594 |
| Natural Place | 1 704 | 254 | 2 568 | 866 | 696 | 340 | 666 |
| Species | 14 247 | 4 893 | 83 359 | - | 7 902 | 38 682 | 36 775 |
| Σ | **949 970** | **301 450** | **8 037 562** | 2 751 105 | 3 437 420 | 536 526 | 1 312 511 |

# 3. Data Fusion

- Next Step: Estimate the quality of the values that can be added to the knowledge base
  - Quality depends on data fusion strategy
- Recent research in this area
  - Knowledge-based Trust as fusion strategy (KBT)
  - Local Closed World Assumption for evaluation (LCWA)

| Country | City |
|---|---|
| Germany | Berlin |
| France | Paris |
| United Kingdom | London |
| Canada | Ottawa |
| USA | Washington D.C. |
| Mexico | Ecatepec |

**Web Table**

✓ ? ✓ ? ✓ ✗

⟺

| Country | Capital |
|---|---|
| Germany | Berlin |
| France | |
| United Kingdom | London |
| Canada | |
| USA | Washington D.C. |
| Mexico | Mexico City |

**Knowledge Base**

Dong, et al.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. KDD 2014.

# Data Fusion – Approach

1. Calculate different trust scores
   1. Baseline: All sources get same score, e.g. 1.0
   2. Knowledge-based Trust: Overlap of values in table and KB
   3. PageRank: PageRank of the web site containing the table
2. Remove values with score below threshold
3. Determine a final value using weighted voting / median

**Germany/Population**

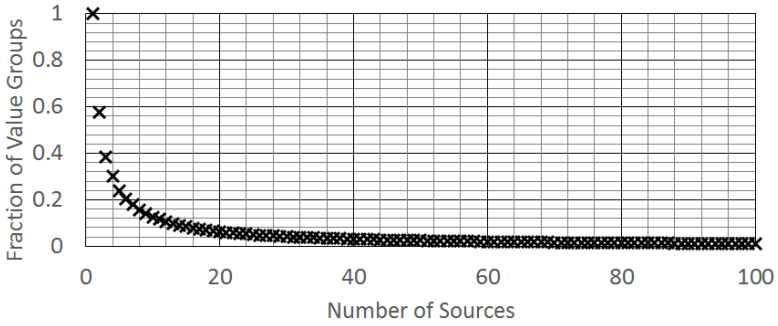
| Source | Value | Score |
|---|---|---|
| A | 8,000,000 | 0.3 |
| B | 81,459,000 | 1.0 |
| C | 81,900,000 | 0.8 |

# Data Fusion – Evaluation

- Evaluation Results

| Strategy | Precision | Recall | F1 |
|---|---|---|---|
| Baseline | .369 | .823 | .509 |
| Knowledge-based Trust | .639 | .785 | .705 |
| PageRank | .365 | .814 | .504 |

- Group Sizes
  - For 58% of all groups we have at least two alternative sources
  - Frequent (Head entities): likely already exist in the KB
  - Infrequent (Tail entities): likely new, but hard to fuse



# Data Fusion – Detailed Results

| DBpedia Class | Existing Values | New Values | Precision | Recall | F1 |
|---|---|---|---|---|---|
| Person | **117 522** | **15 050** | 0.639 | 0.723 | 0.678 |
| Athlete | 84 562 | 9 067 | 0.646 | 0.679 | 0.662 |
| Artist | 2 019 | 427 | 0.711 | 0.830 | 0.766 |
| Office Holder | 3 465 | 510 | 0.698 | 0.849 | 0.766 |
| Politician | 3 124 | 1 167 | 0.533 | 0.765 | 0.628 |
| Organisation | **20 522** | **7 903** | 0.645 | 0.691 | 0.667 |
| Company | 6 376 | 2 547 | 0.700 | 0.834 | 0.761 |
| Sports Team | 790 | 132 | 0.671 | 0.892 | 0.766 |
| Educational Inst. | 8 844 | 3 132 | 0.638 | 0.714 | 0.674 |
| Broadcaster | 4 004 | 1 924 | 0.557 | 0.459 | 0.503 |
| Work | **189 131** | **27 867** | 0.614 | 0.828 | 0.705 |
| Musical Work | 118 511 | 8 427 | 0.599 | 0.830 | 0.695 |
| Film | 29 903 | 12 143 | 0.573 | 0.803 | 0.669 |
| Software | 17 554 | 2 766 | 0.591 | 0.760 | 0.665 |
| Place | **32 855** | **9 871** | 0.767 | 0.858 | 0.810 |
| Populated Place | 16 604 | 6 704 | 0.711 | 0.779 | 0.743 |
| Country | 2 084 | 433 | 0.738 | 0.690 | 0.713 |
| Settlement | 540 | 224 | 0.583 | 0.669 | 0.623 |
| Region | 362 | 70 | 0.587 | 0.784 | 0.671 |
| Architectural Struct. | 10 441 | 1 775 | 0.834 | 0.940 | 0.884 |
| Natural Place | 743 | 64 | 0.843 | 0.940 | 0.889 |
| Species | 9 016 | 1 429 | 0.783 | 0.892 | 0.834 |

# 4. Lessons Learned

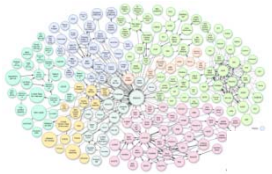
- Web Table data is useful for KB completion, but
- Small number of new values in comparison to overall input
  - Challenge: Improve matching recall, especially for long tail entities
  - Ongoing work:
    - Exploit table context to improve matching
    - Further improve value normalization (surface forms of names, units of measurement) to improve matching
- Knowledge-based trust outperforms other fusion strategies
  - F1 = 0.7 is below the quality required for automating the task
  - Shortcoming: Time dimension not taken into account
  - Ongoing work: Exploit timestamps in webpages and tables

Oulabi, et al.: Fusing Time-Dependent Web Table Data. WebDB 2016.

# Exploiting other Types of Web Data

**schema.org Microdata**

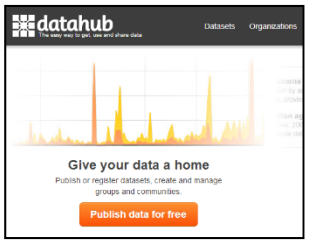

**Linked Data**



**Entity Tables**

| | |
|---|---|
| **Government**[3] | |
| • **Type** | Mayor–Council |
| • **Body** | New York City Council |
| • **Mayor** | Bill de Blasio (D) |
| **Area**[2] | |
| • **Total** | 468.9 sq mi (1,214 km²) |
| • **Land** | 304.8 sq mi (789 km²) |

**Data Portals**




**Public Data Corpora**

- **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/

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

# Thank you!

http://webdatacommons.org/webtables