# Knowledge Base Completion from Web Tables

## An experiment with the Yahoo Knowledge Graph

**DBpedia Meetup 10/27/2016**

Work by Dominique Ritze - Presented by Nicolas Torzec

# About Us...

**Dominique Ritze**

- Ph.D. student, **University of Mannheim, Germany**
- Knowledge base completion from Web tables: **T2K**
- Internship on Slot Filling in the Yahoo Knowledge Graph team

**Nicolas Torzec**

- Lead the Data / Science team for the **Yahoo Knowledge Graph**
- Information extraction, knowledge graph, semantic search, etc.
- NLP+TTS⇒Shopping Comparison⇒Web Search⇒Knowledge Graph

# Yahoo Knowledge Graph

## Large unified knowledge base

- millions of entities, billions of facts
- power knowledge-based services at Yahoo

## Knowledge Acquisition

- extract info about entities from feeds/web
- type inference, schema matching, normalization

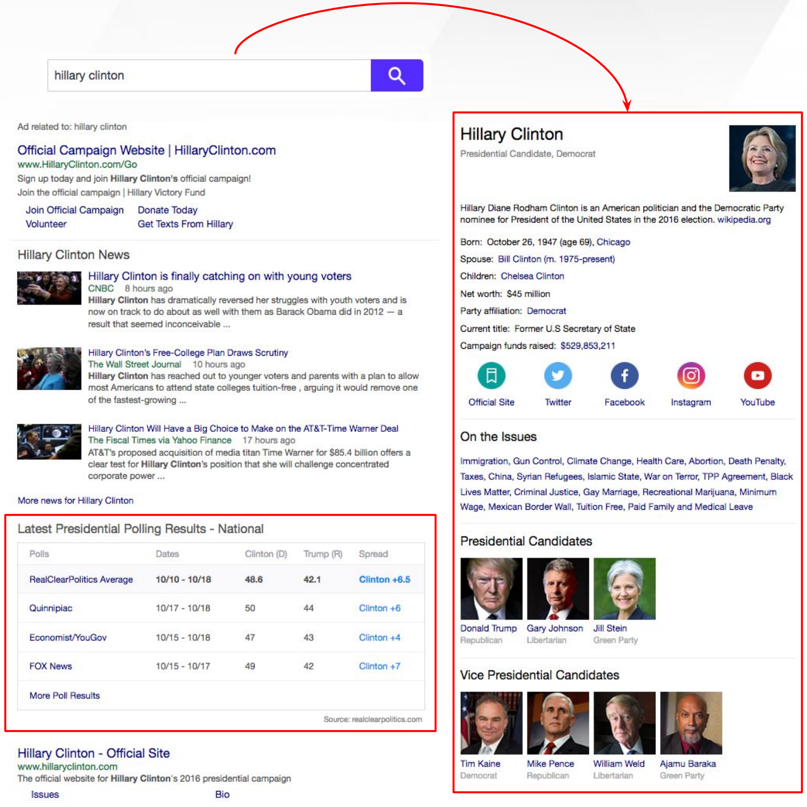
## Knowledge Integration

- cluster entities that refer to same real-world entity
- blend/fuse entity clusters: their types and facts

## Knowledge Enrichments

- notableFor, related entities, hero images
- editorial curation!

## Serving...



hillary clinton

Ad related to: hillary clinton

Official Campaign Website | HillaryClinton.com
www.HillaryClinton.com/Go
Sign up today and join **Hillary Clinton's** official campaign!
Join the official campaign | Hillary Victory Fund

Join Official Campaign | Donate Today
Volunteer | Get Texts From Hillary

Hillary Clinton News

Hillary Clinton is finally catching on with young voters
CNBC 8 hours ago
**Hillary Clinton** has dramatically reversed her struggles with youth voters and is now on track to do about as well with them as Barack Obama did in 2012 — a result that seemed inconceivable ...

Hillary Clinton's Free-College Plan Draws Scrutiny
The Wall Street Journal 10 hours ago
**Hillary Clinton** has reached out to younger voters and parents with a plan to allow most Americans to attend state colleges tuition-free , arguing it would remove one of the fastest-growing ...

Hillary Clinton Will Have a Big Choice to Make on the AT&T-Time Warner Deal
The Fiscal Times via Yahoo Finance 17 hours ago
AT&T's proposed acquisition of media titan Time Warner for $85.4 billion offers a clear test for **Hillary Clinton**'s position that she will challenge concentrated corporate power ...

More news for Hillary Clinton

Latest Presidential Polling Results - National

| Polls | Dates | Clinton (D) | Trump (R) | Spread |
|---|---|---|---|---|
| RealClearPolitics Average | **10/10 - 10/18** | **48.6** | **42.1** | **Clinton +6.5** |
| Quinnipiac | 10/17 - 10/18 | 50 | 44 | Clinton +6 |
| Economist/YouGov | 10/15 - 10/18 | 47 | 43 | Clinton +4 |
| FOX News | 10/15 - 10/17 | 49 | 42 | Clinton +7 |
| More Poll Results | | | | |

Source: realclearpolitics.com

Hillary Clinton - Official Site
www.hillaryclinton.com
The official website for **Hillary Clinton**'s 2016 presidential campaign
Issues | Bio

Hillary Clinton
Presidential Candidate, Democrat

Hillary Diane Rodham Clinton is an American politician and the Democratic Party nominee for President of the United States in the 2016 election. wikipedia.org

Born: October 26, 1947 (age 69), Chicago
Spouse: Bill Clinton (m. 1975-present)
Children: Chelsea Clinton
Net worth: $45 million
Party affiliation: Democrat
Current title: Former U.S Secretary of State
Campaign funds raised: $529,853,211

On the Issues

Immigration, Gun Control, Climate Change, Health Care, Abortion, Death Penalty, Taxes, China, Syrian Refugees, Islamic State, War on Terror, TPP Agreement, Black Lives Matter, Criminal Justice, Gay Marriage, Recreational Marijuana, Minimum Wage, Mexican Border Wall, Tuition Free, Paid Family and Medical Leave

Presidential Candidates

Donald Trump, Republican
Gary Johnson, Libertarian
Jill Stein, Green Party

Vice Presidential Candidates

Tim Kaine, Democrat
Mike Pence, Republican
William Weld, Libertarian
Ajamu Baraka, Green Party

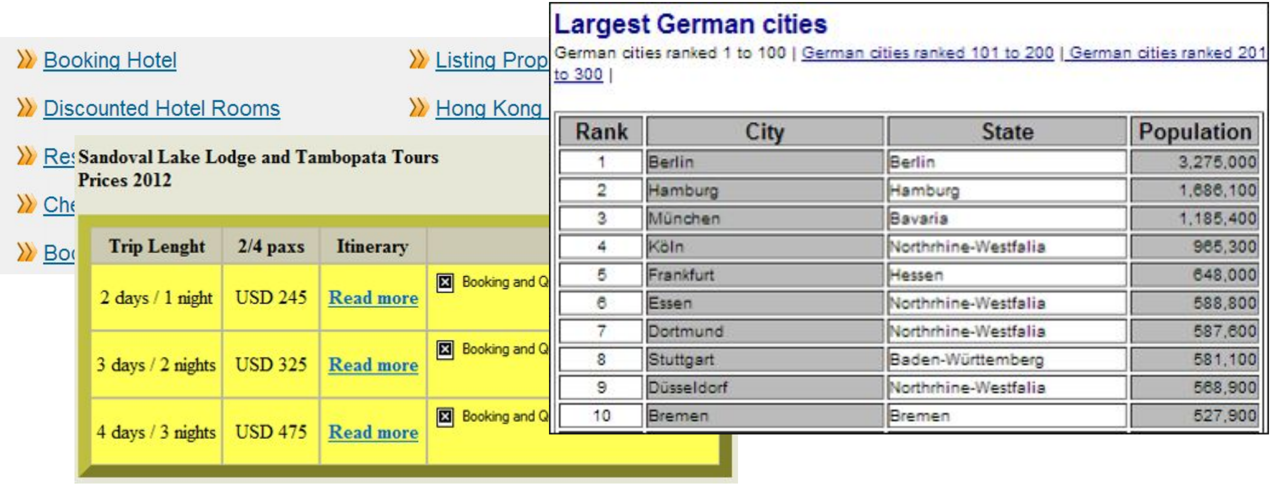
# Knowledge Base Completion from Web Tables

## Goal: enrich KG with information extracted from tables on the Web

- i.e. some information are only (or most easily) available in Web tables.
- **Slot filling** | Set expansion | Attribute discovery | Source discovery | etc.

```
<!DOCTYPE html>
<html>
<table>

</table>
</html>
```



» Booking Hotel

» Listing Prop

» Discounted Hotel Rooms

» Hong Kong

» Res

» Che

» Boo

Sandoval Lake Lodge and Tambopata Tours Prices 2012

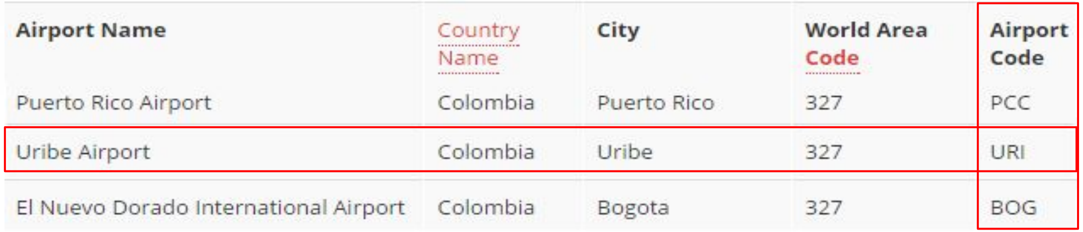
| Trip Lenght | 2/4 paxs | Itinerary | |
|---|---|---|---|
| 2 days / 1 night | USD 245 | Read more | ☒ Booking and Q |
| 3 days / 2 nights | USD 325 | Read more | ☒ Booking and Q |
| 4 days / 3 nights | USD 475 | Read more | ☒ Booking and Q |

**Largest German cities**

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

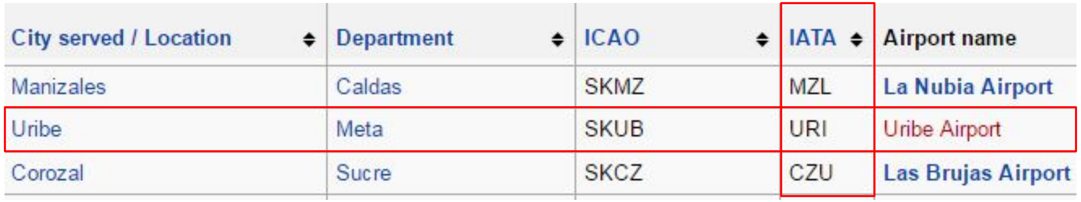
| Rank | City | State | Population |
|---|---|---|---|
| 1 | Berlin | Berlin | 3,275,000 |
| 2 | Hamburg | Hamburg | 1,686,100 |
| 3 | München | Bavaria | 1,185,400 |
| 4 | Köln | Northrhine-Westfalia | 965,300 |
| 5 | Frankfurt | Hessen | 648,000 |
| 6 | Essen | Northrhine-Westfalia | 588,800 |
| 7 | Dortmund | Northrhine-Westfalia | 587,600 |
| 8 | Stuttgart | Baden-Württemberg | 581,100 |
| 9 | Düsseldorf | Northrhine-Westfalia | 568,900 |
| 10 | Bremen | Bremen | 527,900 |

# Knowledge Base Completion from Web Tables (2)



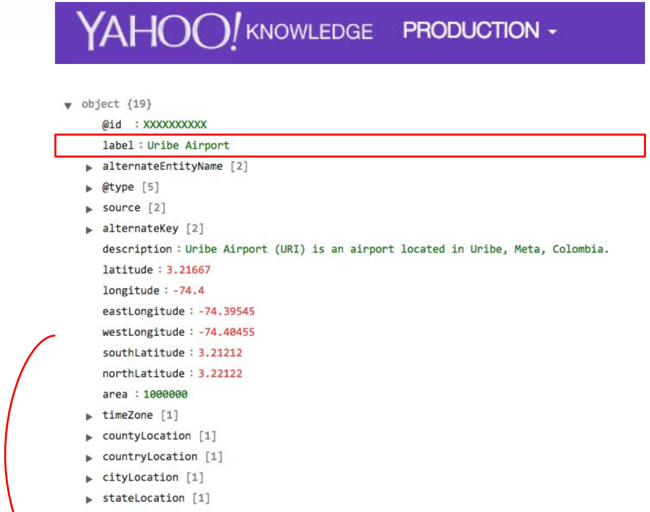
| Airport Name | Country Name | City | World Area Code | Airport Code |
|---|---|---|---|---|
| Puerto Rico Airport | Colombia | Puerto Rico | 327 | PCC |
| Uribe Airport | Colombia | Uribe | 327 | URI |
| El Nuevo Dorado International Airport | Colombia | Bogota | 327 | BOG |

http://www.altiusdirectory.com/Travel/colombia-airports.php



| City served / Location | Department | ICAO | IATA | Airport name |
|---|---|---|---|---|
| Manizales | Caldas | SKMZ | MZL | La Nubia Airport |
| Uribe | Meta | SKUB | URI | Uribe Airport |
| Corozal | Sucre | SKCZ | CZU | Las Brujas Airport |

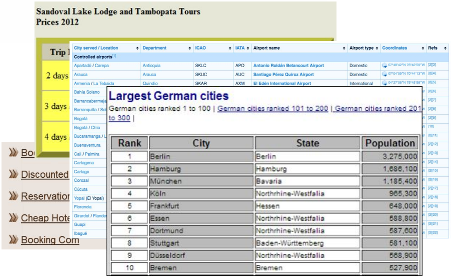
https://en.wikipedia.org/wiki/List_of_airports_in_Colombia



```
▼ object {19}
    @id  : XXXXXXXXXX
    label : Uribe Airport
  ▶ alternateEntityName [2]
  ▶ @type [5]
  ▶ source [2]
  ▶ alternateKey [2]
    description : Uribe Airport (URI) is an airport located in Uribe, Meta, Colombia.
    latitude : 3.21667
    longitude : -74.4
    eastLongitude : -74.39545
    westLongitude : -74.40455
    southLatitude : 3.21212
    northLatitude : 3.22122
    area : 1000000
  ▶ timeZone [1]
  ▶ countyLocation [1]
  ▶ countryLocation [1]
  ▶ cityLocation [1]
  ▶ stateLocation [1]
```

No IATA code for this airport...

# Overview



Sandoval Lake Lodge and Tambopata Tours
Prices 2012

| Trip |
|---|
| 2 days |
| 3 days |
| 4 days |

Boo
Discounted
Reservation
Cheap Hote
Booking Com

| City served / Location | Department | ICAO | IATA | Airport name | Airport type | Coordinates | Refs |
|---|---|---|---|---|---|---|---|
| Controlled airports[1] | | | | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport | Domestic | 07°48′42″N 76°42′59″W | [2][3] |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport | Domestic | 07°04′09″N 70°44′13″W | [2][4] |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport | International | 04°27′06″N 75°45′58″W | [2][5] |
| Bahía Solano | | | | | | | [2][6] |
| Barrancabermeja | | | | | | | [2][7] |
| Barranquilla / Sol | | | | | | | [2][8] |
| Bogotá | | | | | | | [2][9] |
| Bogotá / Chía | | | | | | | [10] |
| Bucaramanga / L | | | | | | | [2][11] |
| Buenaventura | | | | | | | [2][12] |
| Cali / Palmira | | | | | | | [2][13] |
| Cartagena | | | | | | | [2][14] |
| Cartago | | | | | | | [2][15] |
| Corozal | | | | | | | [2][16] |
| Cúcuta | | | | | | | [2][17] |
| Yopal (El Yopal) | | | | | | | [2][18] |
| Florencia | | | | | | | [2][19] |
| Girardot / Flande | | | | | | | [2][20] |
| Guapi | | | | | | | [2][21] |
| Ibagué | | | | | | | [2][22] |

**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 |
| 6 | Essen | Northrhine-Westfalia | 588,800 |
| 7 | Dortmund | Northrhine-Westfalia | 587,600 |
| 8 | Stuttgart | Baden-Württemberg | 581,100 |
| 9 | Düsseldorf | Northrhine-Westfalia | 568,900 |
| 10 | Bremen | Bremen | 527,900 |

**Extraction**

Extract relational tables

(from Web crawl)

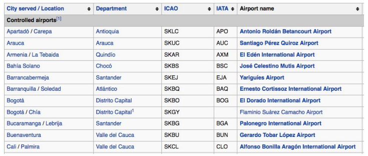
**Knowledge Base**

**Matching**

Match table rows/columns
to KG entities/properties

(type|schema|instance matching)

**Slotting**



| City served / Location | Department | ICAO | IATA | Airport name |
|---|---|---|---|---|
| Controlled airports[1] | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport |
| Bahía Solano | Chocó | SKBS | BSC | José Celestino Mutis Airport |
| Barrancabermeja | Santander | SKEJ | EJA | Yariguíes Airport |
| Barranquilla / Soledad | Atlántico | SKBQ | BAQ | Ernesto Cortissoz International Airport |
| Bogotá | Distrito Capital | SKBO | BOG | El Dorado International Airport |
| Bogotá / Chía | Distrito Capital[1] | SKGY | | Flaminio Suárez Camacho Airport |
| Bucaramanga / Lebrija | Santander | SKBG | BGA | Palonegro International Airport |
| Buenaventura | Valle del Cauca | SKBU | BUN | Gerardo Tobar López Airport |
| Cali / Palmira | Valle del Cauca | SKCL | CLO | Alfonso Bonilla Aragón International Airport |

**Fusion**

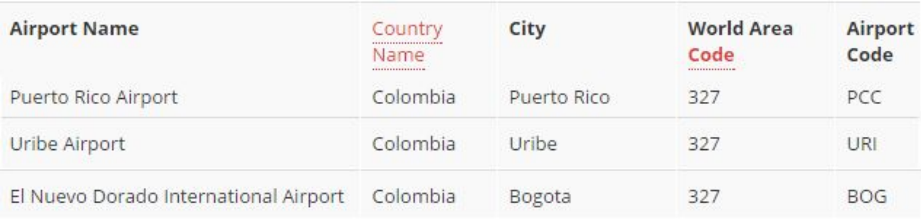
Deduplicate and identify most likely
value of each table entity/property

(across tables)

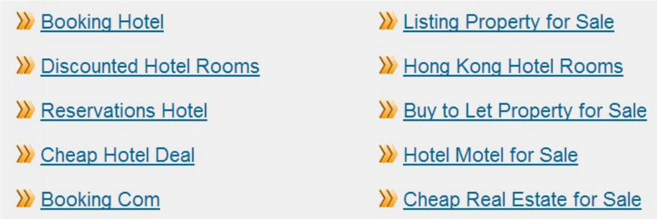
# Table Extraction

## Goal: extract relational tables

- e.g. layout tables vs. relational tables vs. single-entity tables vs. ...



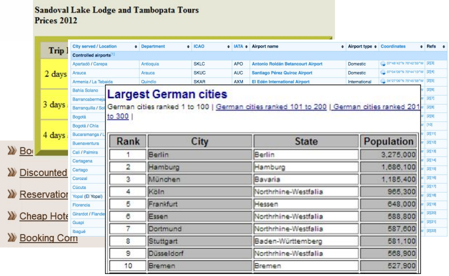
| Airport Name | Country Name | City | World Area Code | Airport Code |
|---|---|---|---|---|
| Puerto Rico Airport | Colombia | Puerto Rico | 327 | PCC |
| Uribe Airport | Colombia | Uribe | 327 | URI |
| El Nuevo Dorado International Airport | Colombia | Bogota | 327 | BOG |



- » Booking Hotel
- » Discounted Hotel Rooms
- » Reservations Hotel
- » Cheap Hotel Deal
- » Booking Com
- » Listing Property for Sale
- » Hong Kong Hotel Rooms
- » Buy to Let Property for Sale
- » Hotel Motel for Sale
- » Cheap Real Estate for Sale

## Classification problem

- i.e. distinguish relational tables from other tables
- Features: level of nesting, avg cell length, % of links, datatype heterogeneity, etc.
- 11B tables in WDC Common Crawl 2012 ⇒ **1.3% are relational** (~91M tables)



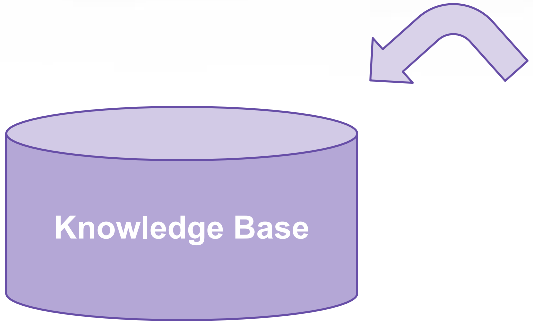
Sandoval Lake Lodge and Tambopata Tours
Prices 2012

| City served / Location | Department | ICAO | IATA | Airport name | Airport type | Coordinates | Refs |
|---|---|---|---|---|---|---|---|
| Controlled airports[1] | | | | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport | Domestic | 07°48′42″N 76°42′59″W | [2][3] |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport | Domestic | 07°04′09″N 70°44′13″W | [2][4] |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport | International | 04°27′06″N 75°45′58″W | [2][5] |

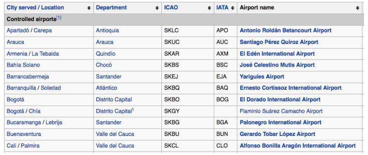
## Largest German cities

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 |
| 6 | Essen | Northrhine-Westfalia | 588,800 |
| 7 | Dortmund | Northrhine-Westfalia | 587,600 |
| 8 | Stuttgart | Baden-Württemberg | 581,100 |
| 9 | Düsseldorf | Northrhine-Westfalia | 568,900 |
| 10 | Bremen | Bremen | 527,900 |



# Slotting



| City served / Location | Department | ICAO | IATA | Airport name |
|---|---|---|---|---|
| Controlled airports[1] | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport |
| Bahía Solano | Chocó | SKBS | BSC | José Celestino Mutis Airport |
| Barrancabermeja | Santander | SKEJ | EJA | Yariguíes Airport |
| Barranquilla / Soledad | Atlántico | SKBQ | BAQ | Ernesto Cortissoz International Airport |
| Bogotá | Distrito Capital | SKBO | BOG | El Dorado International Airport |
| Bogotá / Chía | Distrito Capital[1] | SKGY | | Flaminio Suárez Camacho Airport |
| Bucaramanga / Lebrija | Santander | SKBG | BGA | Palonegro International Airport |
| Buenaventura | Valle del Cauca | SKBU | BUN | Gerardo Tobar López Airport |
| Cali / Palmira | Valle del Cauca | SKCL | CLO | Alfonso Bonilla Aragón International Airport |

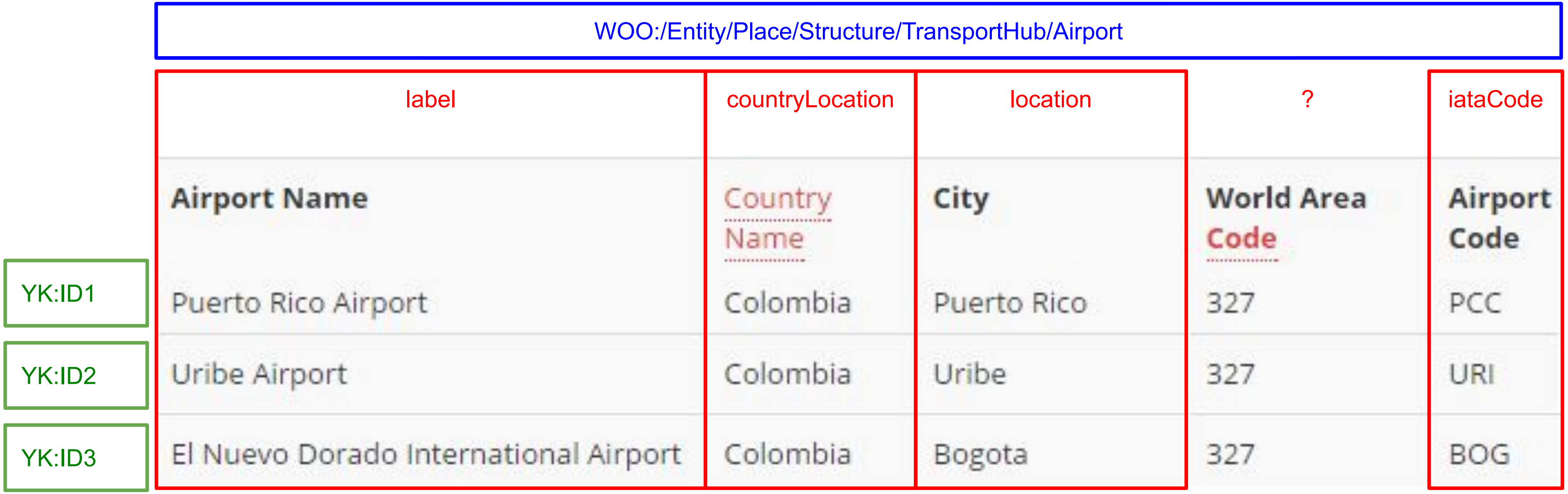
# Extraction

# Matching

# Fusion

# Matching

**Goal = match table/rows/columns to type, entities, and properties**



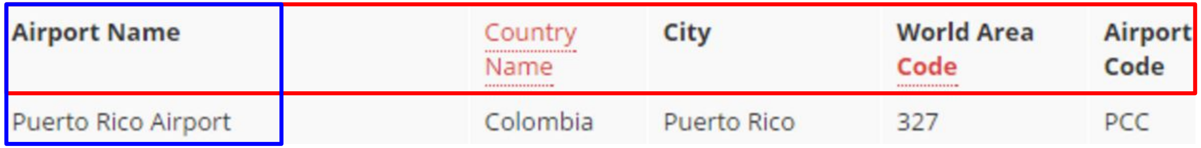
WOO:/Entity/Place/Structure/TransportHub/Airport

| | label | countryLocation | location | ? | iataCode |
|---|---|---|---|---|---|
| | **Airport Name** | **Country Name** | **City** | **World Area Code** | **Airport Code** |
| YK:ID1 | Puerto Rico Airport | Colombia | Puerto Rico | 327 | PCC |
| YK:ID2 | Uribe Airport | Colombia | Uribe | 327 | URI |
| YK:ID3 | El Nuevo Dorado International Airport | Colombia | Bogota | 327 | BOG |

# Matching: Preprocessing

## Table preprocessing

- Filter low-quality tables (domain, page, language, number of rows/cells)
- Standardize cells (datatype casting, unit conversion)
- Detect header row, key column (based on datatype, uniqueness, position)
- ⇒ **Subset of 30M tables** (589M rows / 2.6B cells)



| Airport Name | | Country Name | City | World Area Code | Airport Code |
|---|---|---|---|---|---|
| Puerto Rico Airport | | Colombia | Puerto Rico | 327 | PCC |

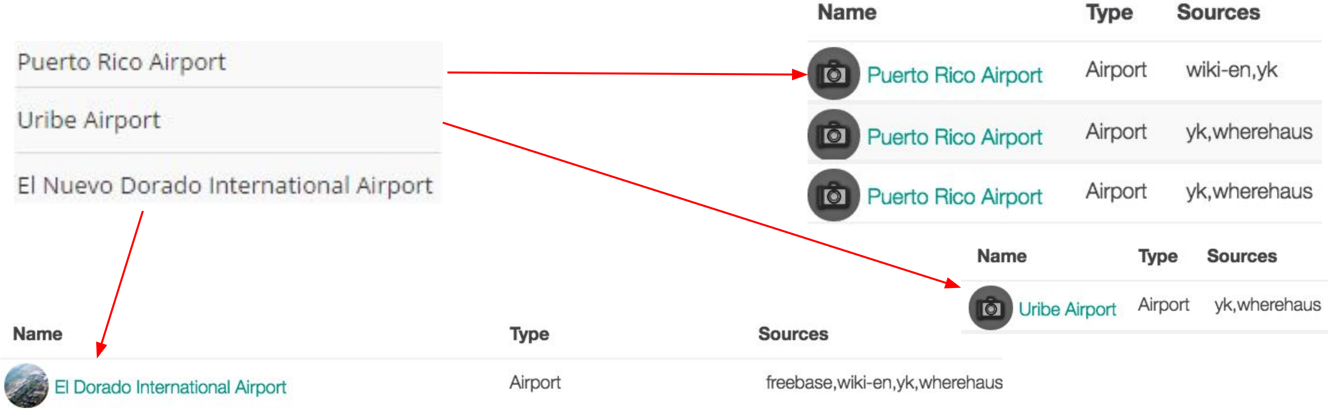
## KB preprocessing

- Filter entities based on provenance, language, type, richness
- Filter properties irrelevant for matching ; denormalize object properties
- ⇒ **Subset of 56M entities** (854M facts) (~10x DBpedia)

# Matching: Blocking (1)

## Identify most likely "KG entities" for each "table entity"

- Why: to prevent full pair-wise comparison (i.e. **589M rows x 56M entities)**
- How: exact match on label/aliases, or multi-attribute fuzzy matching



| |
|---|
| Puerto Rico Airport |
| Uribe Airport |
| El Nuevo Dorado International Airport |

| Name | Type | Sources |
|---|---|---|
| Puerto Rico Airport | Airport | wiki-en,yk |
| Puerto Rico Airport | Airport | yk,wherehaus |
| Puerto Rico Airport | Airport | yk,wherehaus |

| Name | Type | Sources |
|---|---|---|
| Uribe Airport | Airport | yk,wherehaus |

| Name | Type | Sources |
|---|---|---|
| El Dorado International Airport | Airport | freebase,wiki-en,yk,wherehaus |

- **⇒ 130M rows x 21M entities ; median number of matches: 7** (vs. DBpedia: 1)

# Matching: Blocking (2)

## Identify most likely “KG types” for each table

- Why: to limit further the number of comparisons and improve precision.
- Features: types of matched entities, type specificity/relatedness, etc.

| | Airport | Place | Person | MusicTrack |
|---|---|---|---|---|
| Puerto Rico Airport | 3 | 3 | 0 | 1 |
| Uribe Airport | 1 | 1 | 0 | 1 |
| El Nuevo Dorado International Airport | 1 | 1 | 0 | 0 |

## Based upon most likely types:

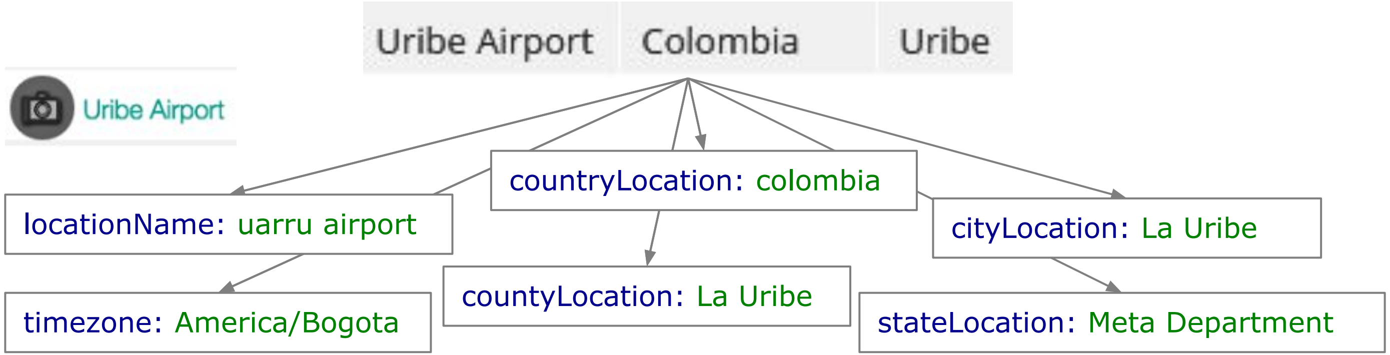
- Filter out rows whose type does not match the type of the table
- Redo entity blocking, focusing on entities whose type match one of the most likely types.

# Matching: Type Inference & Schema Matching (1)

## Assign a KG type to each table and a KG property to each column

1) Compute similarities between table cells and entity facts

- Datatype-specific comparisons
- Set comparisons



| | | |
|---|---|---|
| Uribe Airport | Colombia | Uribe |

Uribe Airport

locationName: uarru airport

countryLocation: colombia

cityLocation: La Uribe

timezone: America/Bogota

countyLocation: La Uribe

stateLocation: Meta Department

# Matching: Type Inference & Schema Matching (2)

## Assign a KG type to each table and a KG property to each column

2) Infer column-to-property mapping

- Naive approach: aggregate cell/fact similarity at the column level (⇒ ML)

3) infer table to type mapping

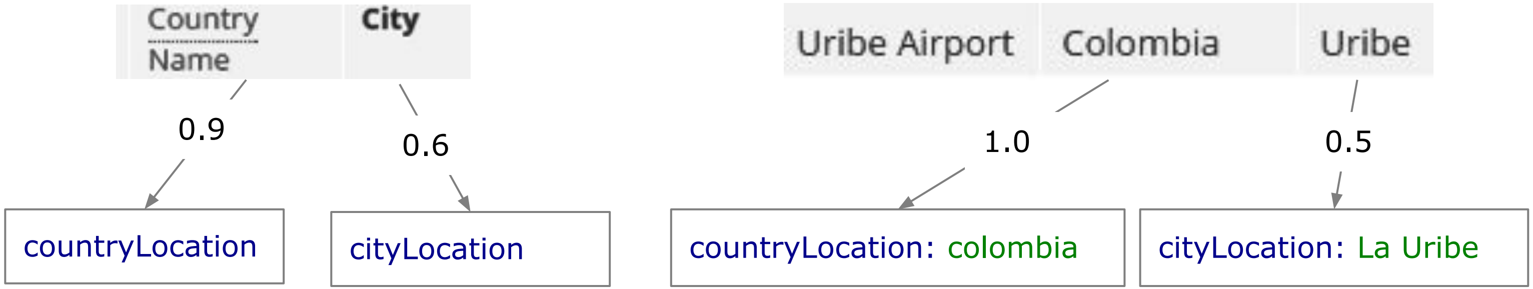
- Naive approach: aggregate column/property similarity at the table level (⇒ ML)

| Column: city | Puerto Rico Airport | | | Uribe Airport | El Nuevo Dorado Int. Airport | |
|---|---|---|---|---|---|---|
| **KG property** | **YK1** | **YK2** | **YK3** | **YK4** | **YK5** | **score** |
| cityLocation | 1.0 | 0.8 | 1.0 | 0.5 | 0.7 | 4.0 |
| countryLocation | 0.0 | 0.2 | 0.0 | 0.5 | 0.1 | 0.8 |

# Matching: Instance Matching

## Match each table entity to a KG entity

- Classification problem
- Features: property similarity score, value similarity score, entity type, etc.



Country Name

City

0.9

0.6

countryLocation

cityLocation

Uribe Airport

Colombia

Uribe

1.0

0.5

countryLocation: colombia

cityLocation: La Uribe

*naive similarity score ⇒ 0.9*1.0 + 0.6*0.5*

- **Naive classifier ⇒ ML classifier**

# Matching: Some Results

## Naive preprocessing / blocking

- Tables: from 91M tables to 30M tables (589M rows x 2.6B cells)
- KG: subset of 56M entities x 854M facts (~10x DBpedia)
- Blocking ⇒ 130M rows x 21M entities (median number of matches per row: 7)

## Naive matching

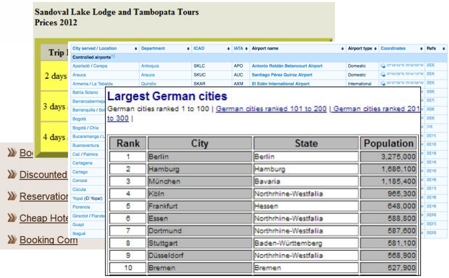
- table2type: 7M correspondences
- column2property: 9M correspondences
- row2entity: 82M correspondences

Room for improvement: new features, ML, joint optimization

- Number of generated triples: 144M
- Number of unique value across entity-property group: 13M
- ~16% of the groups do not overlap with YK

Need data **fusion** to find most likely value for each group

## Slotting



Sandoval Lake Lodge and Tambopata Tours
Prices 2012

| Trip |
|---|
| 2 days |
| 3 days |
| 4 days |

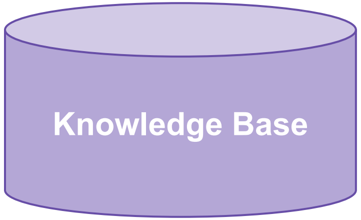
» Boo
» Discounted
» Reservatior
» Cheap Hote
» Booking Com

| City served / Location | Department | ICAO | IATA | Airport name | Airport type | Coordinates | Refs |
|---|---|---|---|---|---|---|---|
| Controlled airports[1] | | | | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport | Domestic | 07°48′42″N 76°42′59″W | [2][3] |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport | Domestic | 07°04′09″N 70°44′13″W | [2][4] |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport | International | 04°27′06″N 75°45′58″W | [2][5] |
| Bahía Solano | | | | | | | [2][6] |
| Barrancabermeja | | | | | | | [2][7] |
| Barranquilla / Sol | | | | | | | [2][8] |
| Bogotá | | | | | | | [2][9] |
| Bogotá / Chía | | | | | | | [10] |
| Bucaramanga / L | | | | | | | [2][11] |
| Buenaventura | | | | | | | [2][12] |
| Cali / Palmira | | | | | | | [2][13] |
| Cartagena | | | | | | | [2][14] |
| Cartago | | | | | | | [2][15] |
| Corozal | | | | | | | [2][16] |
| Cúcuta | | | | | | | [2][17] |
| Yopal (El Yopal) | | | | | | | [2][18] |
| Florencia | | | | | | | [2][19] |
| Girardot / Flande | | | | | | | [2][20] |
| Guapi | | | | | | | [2][21] |
| Ibagué | | | | | | | [2][22] |

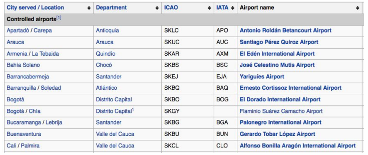
### Largest German cities

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 |
| 6 | Essen | Northrhine-Westfalia | 588,800 |
| 7 | Dortmund | Northrhine-Westfalia | 587,600 |
| 8 | Stuttgart | Baden-Württemberg | 581,100 |
| 9 | Düsseldorf | Northrhine-Westfalia | 568,900 |
| 10 | Bremen | Bremen | 527,900 |



Knowledge Base



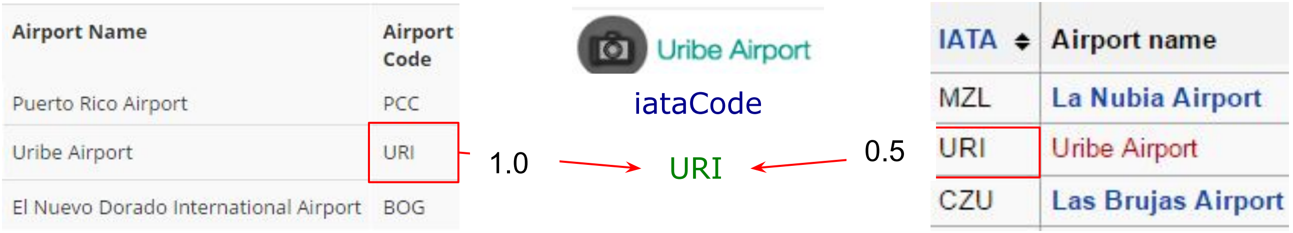
| City served / Location | Department | ICAO | IATA | Airport name |
|---|---|---|---|---|
| Controlled airports[1] | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport |
| Bahía Solano | Chocó | SKBS | BSC | José Celestino Mutis Airport |
| Barrancabermeja | Santander | SKEJ | EJA | Yariguíes Airport |
| Barranquilla / Soledad | Atlántico | SKBQ | BAQ | Ernesto Cortissoz International Airport |
| Bogotá | Distrito Capital | SKBO | BOG | El Dorado International Airport |
| Bogotá / Chía | Distrito Capital[1] | SKGY | | Flaminio Suárez Camacho Airport |
| Bucaramanga / Lebrija | Santander | SKBG | BGA | Palonegro International Airport |
| Buenaventura | Valle del Cauca | SKBU | BUN | Gerardo Tobar López Airport |
| Cali / Palmira | Valle del Cauca | SKCL | CLO | Alfonso Bonilla Aragón International Airport |

## Extraction → Matching → Fusion

# Fusion

## Goal: identify most likely value for each table entity/property group

- Local Closed World Assumption (LCWA): everything in the KB is true
- Knowledge Base Trust (KBT) to assign a reliability score to a column
- Filter out not reliable columns
- “Most likely value” = weighted majority/median (with KBT score as weight)



| Airport Name | Airport Code |
|---|---|
| Puerto Rico Airport | PCC |
| Uribe Airport | URI |
| El Nuevo Dorado International Airport | BOG |

| IATA | Airport name |
|---|---|
| MZL | La Nubia Airport |
| URI | Uribe Airport |
| CZU | Las Brujas Airport |

- Works best for large tables and facts coming from many different tables!

# Fusion Some Results

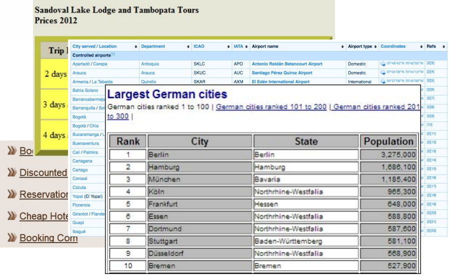
## Fusion results (with naive approach)

- Metrics: precision/recall based upon overlapping facts
- Slight deviations allowed...

| | Missing facts | Overlapping facts | Precision |
|---|---|---|---|
| MediaRelease.albumTrackOf | 65,867 | 105,484 | 0.92 |
| EducationalInstitution.countryLocation | 1336 | 1313 | 0.99 |
| CreativeWork.releaseDate | 0 | 208,626 | 0.74 |

## Expected slotting (with naive approach)

- 190K new/missing facts with exp. precision of 0.99 (2% of all facts)
- 480K new/missing facts with exp. precision of 0.86 (4% of all facts)



Sandoval Lake Lodge and Tambopata Tours
Prices 2012

## Largest German cities

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

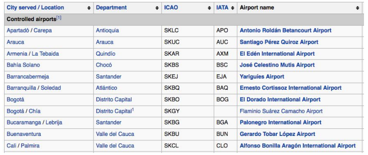
| Rank | City | State | Population |
|---|---|---|---|
| 1 | Berlin | Berlin | 3,275,000 |
| 2 | Hamburg | Hamburg | 1,686,100 |
| 3 | München | Bavaria | 1,185,400 |
| 4 | Köln | Northrhine-Westfalia | 965,300 |
| 5 | Frankfurt | Hessen | 648,000 |
| 6 | Essen | Northrhine-Westfalia | 588,800 |
| 7 | Dortmund | Northrhine-Westfalia | 587,600 |
| 8 | Stuttgart | Baden-Württemberg | 581,100 |
| 9 | Düsseldorf | Northrhine-Westfalia | 568,900 |
| 10 | Bremen | Bremen | 527,900 |

# Extraction

Knowledge Base

# Matching

# Slotting



| City served / Location | Department | ICAO | IATA | Airport name |
|---|---|---|---|---|
| Controlled airports[1] | | | | |
| Apartadó / Carepa | Antioquia | SKLC | APO | Antonio Roldán Betancourt Airport |
| Arauca | Arauca | SKUC | AUC | Santiago Pérez Quiroz Airport |
| Armenia / La Tebaida | Quindío | SKAR | AXM | El Edén International Airport |
| Bahía Solano | Chocó | SKBS | BSC | José Celestino Mutis Airport |
| Barrancabermeja | Santander | SKEJ | EJA | Yariguíes Airport |
| Barranquilla / Soledad | Atlántico | SKBQ | BAQ | Ernesto Cortissoz International Airport |
| Bogotá | Distrito Capital | SKBO | BOG | El Dorado International Airport |
| Bogotá / Chía | Distrito Capital[1] | SKGY | | Flaminio Suárez Camacho Airport |
| Bucaramanga / Lebrija | Santander | SKBG | BGA | Palonegro International Airport |
| Buenaventura | Valle del Cauca | SKBU | BUN | Gerardo Tobar López Airport |
| Cali / Palmira | Valle del Cauca | SKCL | CLO | Alfonso Bonilla Aragón International Airport |

# Fusion

# Take Aways

- Large amounts of facts available in Web tables
- ⇒ Knowledge Base Completion from Web tables

- Naive end-to-end system running on Yahoo Knowledge Graph
- Slot filling: 480K new facts @ expected precision of 0.86
- Decent baseline but improvements are possible at each step!

- Future directions:
  - ML and KG as feedback loop
  - Joint optimization problem
  - Topic/Domain-specific applications

# Thank You.

torzecn@yahoo-inc.com
twitter: @nicolastorzec

yes, we are hiring too.