Schema.org Annotations and Web Tables

Underexploited Semantic Nuggets on the Web?

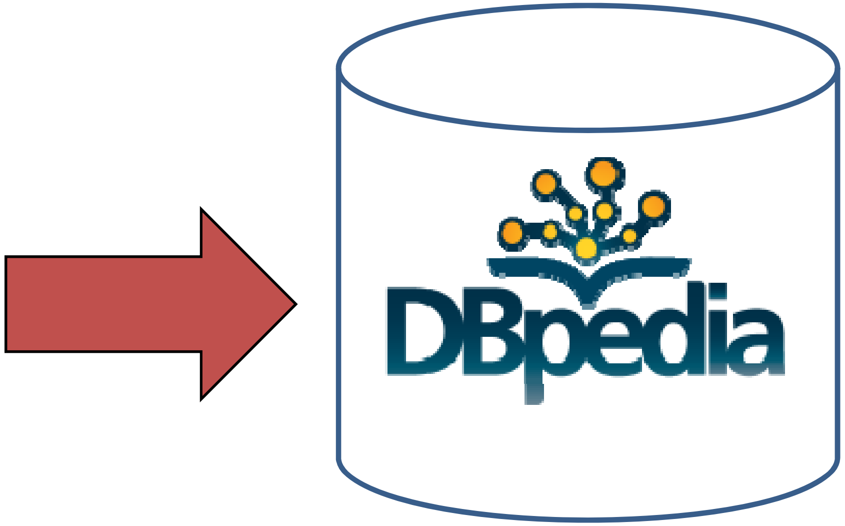
UNIVERSITY OF MANNHEIM
Data and Web Science Group

Prof. Dr. Christian Bizer

LANGUAGE, DATA and KNOWLEDGE 2019

20-23 May 2019 in Leipzig, Germany

Cool Resources are often Build Exploiting Some Structure Somewhere



Web Tables

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		
4	Cologne, North Rhine-Westphalia		
5	Frankfurt am Main, Hesse		
6			
7			
8			
9			
10			

NAME	WEBSITE		
Aeroflot	www.aeroflot.ru/e		
Air France	www.airfrance.fr		
All Nippon Airways	www.ana.co.jp		
Asiana Airlines	www.flyasiana.co		
Cathay Pacific	www.cathaypacifi		
China Airlines	www.china-airlines.com	Taipei	CI
			CZ
			JL

150 INTERNATIONAL AFFAIRS		2016	2017	2018	2019	2020
8.	End Funding for the United Nations Development Program (UNDP)	81	81	82	83	85
9.	End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)	10	10	10	10	11
10.	Eliminate the U.S. Trade and Development Agency (USTDA)	56	56	56	57	58
11.	Reform Food Aid Programs	168	168	168	170	173
12.	Eliminate Export-Import Bank	200	200	200	200	200
13.	Eliminate the Overseas Priva Investment Corporation (OP					
14.	Eliminate Funding for the Un Nations Population Fund (UN					
SUBTOTALS						

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

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	Cupid	Cupid Shuffle
6	Bon Jovi	Livin' On A Prayer
7	Beyonce	Single Ladies (Put A Ring On It)
8	Diamond, Neil	Sweet Caroline (Good Times Never Seemed So Good)
9	Morrison, Van	Brown Eyed Girl
10	Def Leppard	Pour Some Sugar On Me
11	B-52's	Love Shack
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

Schema.org Annotations

UNIVERSITY OF MANNHEIM
Data and Web Science Group

```
<div  itemtype="http://schema.org/Hotel">
     <span itemprop="name">Vienna Marriott Hotel</span>
     <span itemprop="address" itemtype="http://schema.org/PostalAddress">
          <span itemprop="streetAddress">Parkring 12a</span>
          <span itemprop="addressLocality">Vienna</span>
     </span>
</div>
```

```
<div  itemtype="http://schema.org/Product">
     <h1 itemprop="name">Signature Instant Dome 7</h1>
    Item # <span itemprop="productID">200734246807</span>
    <img itemprop="image" src="../20073424680.jpg">
    <span itemprop="review" itemtype="http://schema.org/Review">
          <span itemprop="ratingValue">5</span>
          <span itemprop="reviewBody">Great waterproof tent!</span>
     </span>
</div>
```

Hypothesis

- Web tables and Schema.org annotations are useful training data for tasks such as
  - Named Entity Recognition
  - Information Extraction
  - Entity Matching
  - Taxonomy Induction
  - Sentiment Analysis
  - Knowledge Base Completion
- but they are hardly used by public research yet

Outline

1. Schema.org Annotations
   - Adoption on the Web
   - Current applications and further potential
2. Web Tables
   - Usage on the Web
   - Application potential
3. Conclusions

1. Schema.org Annotations

schema.org

Search

Home Schemas Documentation

Thing > Organization > LocalBusiness

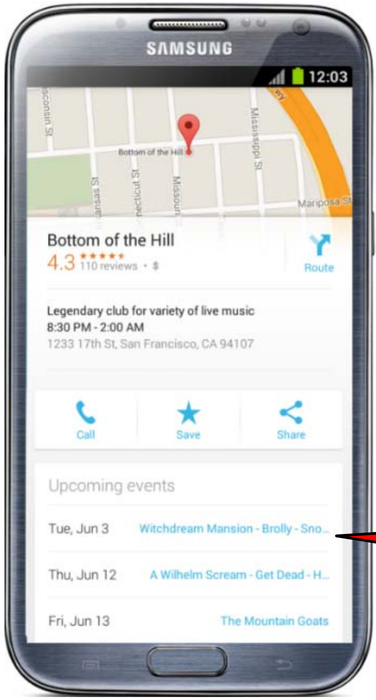
A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.

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

- ask site owners since 2011 to annotate data for enriching search results
- 675 Types: Event, Place, Local Business, Product, Review, Person
- Encoding: Microdata, RDFa, JSON-LD

Usage of Schema.org Data @ Google

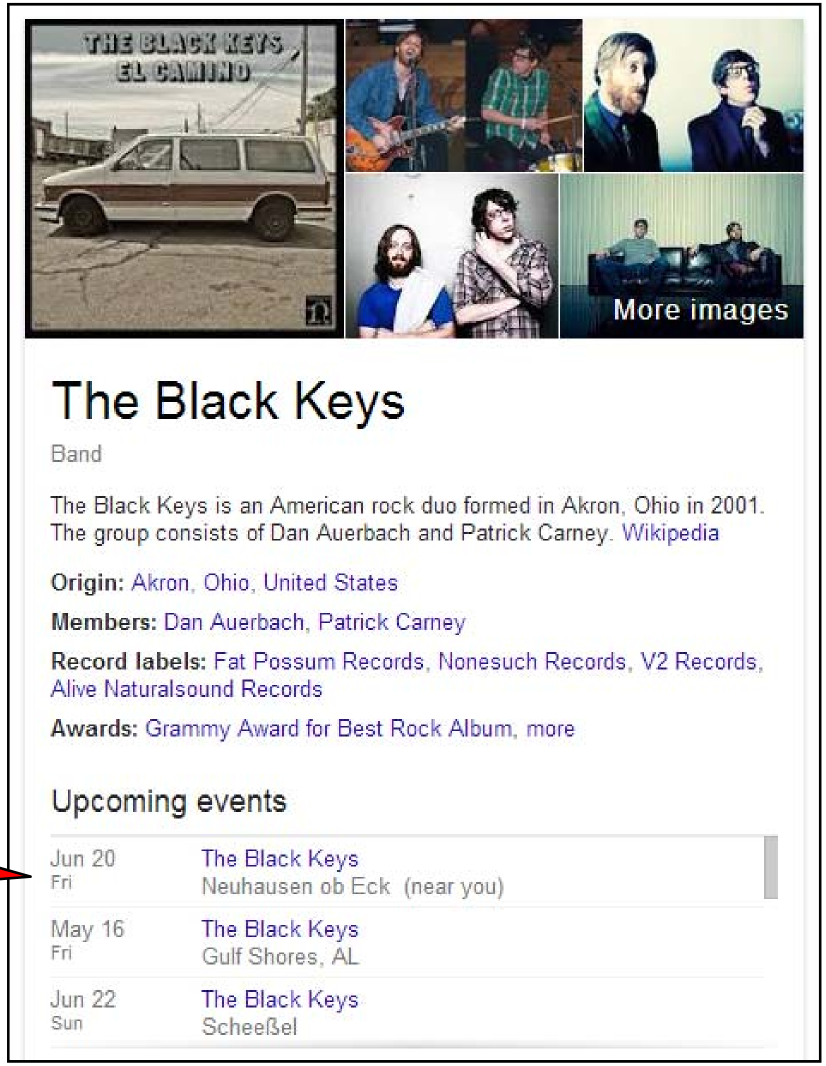
UNIVERSITY OF MANNHEIM
Data and Web Science Group



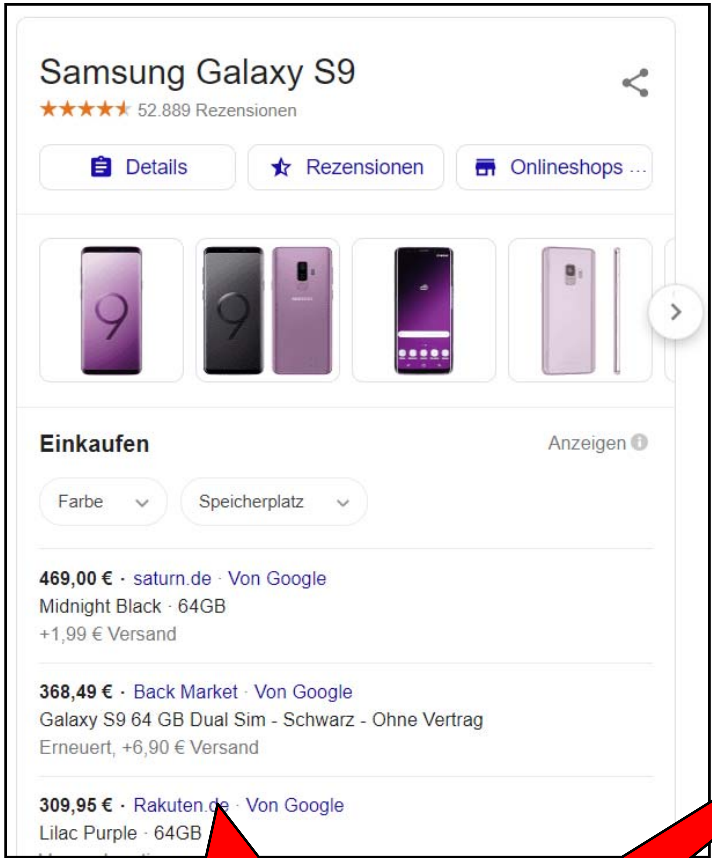
Data snippets within search results

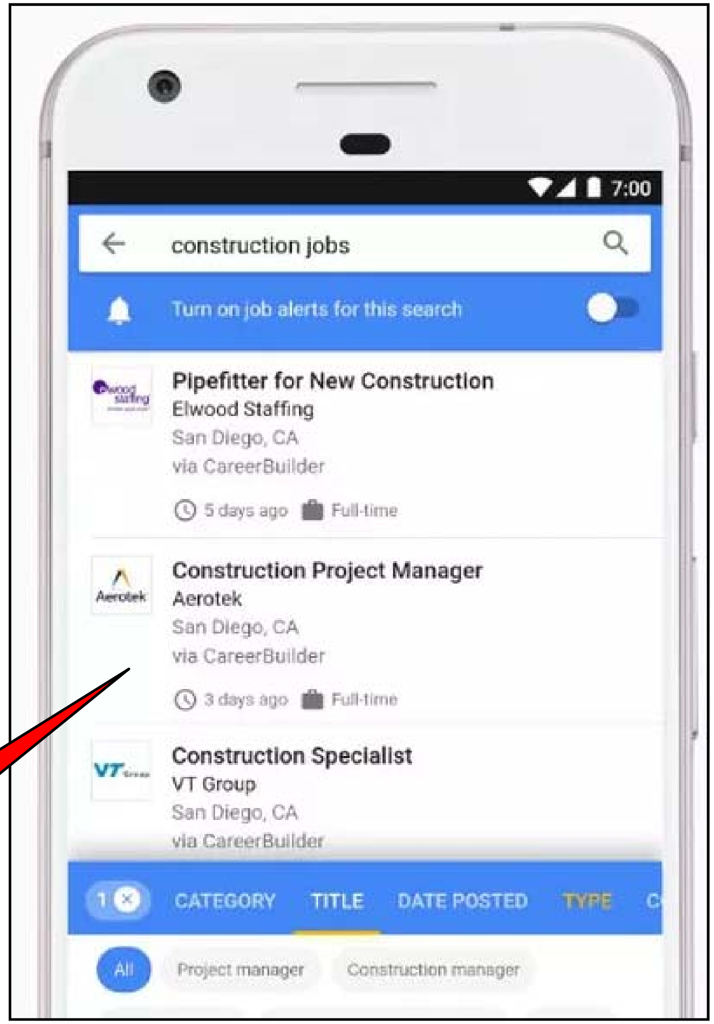
Data snippets within info boxes

Google maps location data



Usage of Schema.org Data @ Google





Product offers

Job offers

https://developers.google.com/search/docs/guides/search-gallery

Web Data Commons Project

- extracts all Microformat, Microdata, RDFa, JSON-LD data from the Common Crawl
- analyzes and provides the extracted data for download
- statistics about some extraction runs
  - 2018 CC Corpus: 2.5 billion HTML pages → 31.5 billion RDF triples
  - 2017 CC Corpus: 3.1 billion HTML pages → 38.2 billion RDF triples
  - 2014 CC Corpus: 2.0 billion HTML pages → 20.4 billion RDF triples
  - 2010 CC Corpus: 2.8 billion HTML pages → 5.1 billion RDF triples
- uses 100 machines on Amazon EC2
  - approx. 2000 machine/hours → 250 Euro

http://webdatacommons.org/structureddata/

Overall Adoption 2018

UNIVERSITY OF MANNHEIM
Data and Web Science Group

944 million HTML pages out of the 2.5 billion pages provide semantic annotations (37.1%).

9.6 million pay-level-domains (PLDs) out of the 32.8 million pay-level-domains covered by the crawl provide semantic annotations (29.3%).

http://webdatacommons.org/structureddata/2018-12/

Frequently used Schema.org Classes

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Top Classes	# Websites (PLDs)	
	Microdata	JSON-LD
schema:WebPage	1,124,583	121,393
schema:Product	812,205	40,169
schema:Offer	676,899	57,756
schema:BreadcrumbList	621,344	205,971
schema:Article	612,361	57,082
schema:Organization	510,069	1,349,775
schema:PostalAddress	502,615	176,500
schema:ImageObject	360,875	111,946
schema:Blog	337,843	12,174
schema:Person	324,349	335,784
schema:LocalBusiness	294,390	249,017
schema:AggregateRating	258,078	23,105
schema:Review	124,022	6,622
schema:Place	92,127	66,396
schema:Event	88,130	63,605

http://webdatacommons.org/structureddata/2018-12/

Attributes used to Describe Products

Top Attributes	PLDs Microdata	
	#	%
schema:Product/name	754,812	92 %
schema:Product/offers	645,994	79 %
schema:Offer/price	639,598	78 %
schema:Offer/priceCurrency	606,990	74 %
schema:Product/image	573,614	70 %
schema:Product/description	520,307	64 %
schema:Offer/availability	477,170	58 %
schema:Product/url	364,889	44 %
schema:Product/sku	160,343	19 %
schema:Product/aggregateRating	141,194	17 %
schema:Product/brand	113,209	13 %
schema:Product/category	62,170	7 %
schema:Product/productID	47,088	5 %
…	…	…

http://webdatacommons.org/structureddata/2018-12/stats/html-md.xlsx

Use Case 1: Gather Large Amounts of Hyponym Relationships

- plenty of entity names are annotated
- Hearst Pattern of the 21st century
  (if class is covered)

Class	# Entities Names
Product	368,386,801
Organization	176,365,220
Local Business	44,100,625
Place	31,748,612
Event	26,035,295
Job Posting	5,078,941
Movie	5,233,520
Hotel	2,537,089
College or University	2,392,797
Recipe	2,036,417
Music Album	1,594,941
Restaurant	1,480,443
Airport	1,145,323

http://webdatacommons.org/structureddata/2018-12/stats/schema_org_subsets.html

Coverage: Top 15 Hotel Websites

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Top 15 Travel Websites	schema:Hotel
Booking.com	✓
TripAdvisor	✓
Expedia	✓
Agoda	✓
Hotels.com	✓
Kayak	✓
Priceline	✓
Travelocity	✓
Orbitz	✓
ChoiceHotels	✓
HolidayCheck	✓
ChoiceHotels	✓
InterContinental Hotels Group	✓
Marriott International	✓
Global Hyatt Corp.	☒

Adoption:
93 %

Kärle, Fensel, Toma, Fensel: **Why Are There More Hotels in Tyrol than in Austria? Analyzing Schema.org Usage in the Hotel Domain.** ICTT 2016.

Use Case 2: Distant Supervision for Information Extraction

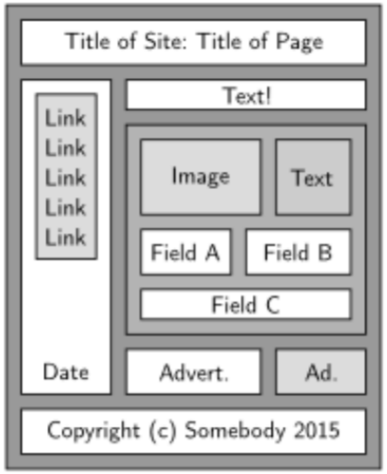
- **Goal:** Learn how to extract data from websites that do not provide schema.org annotations using annotations as training data.
- **Example:** Google Small Events Paper @ SIGIR 2015
  - extract data about small venue concerts, theatre performances, garage sales, etc.
  - they use 217,000 schema.org events from ClueWeb2012 as distant supervision, e.g.

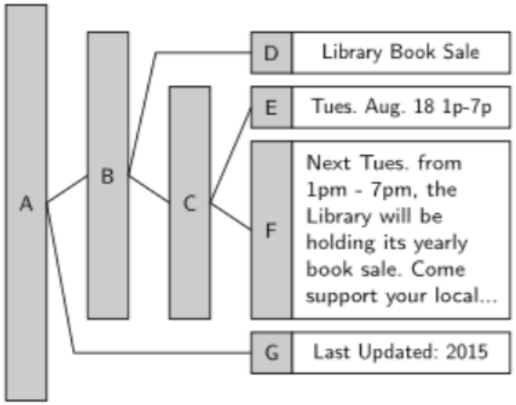
```
<div  itemtype="http://schema.org/MusicEvent">
    <h1 itemprop="name">School Cool Concert</h1>
   <p itemprop="startDate">25.09.2018 15:00</p>
    <p itemprop="location" itemtype="http://schema.org/Place">School Gym Puero High</span>
<div>
```

Foley, Bendersky, Josifovski: **Learning to Extract Local Events from the Web.** SIGIR2015.

Google Small Events Paper @ SIGIR 2015

- **Approach**
  - train an information extraction method that combines structural and text features
  - extracted attributes: What? When? Where?
  - apply method to pages without semantic annotations
- **Result**
  - extract 200,000 additional small-scale events from ClueWeb2012
  - double recall at 0.85 precision





Foley, Bendersky, Josifovski: **Learning to Extract Local Events from the Web.** SIGIR2015.

Underexploited?

University of Mannheim
Data and Web Science Group

- Using schema.org data as distant supervision will likely also work for other information extraction tasks.
- **Research questions**
  - How to best exploit annotations as distant supervision?
  - How good do the annotations perform compared to other sources of supervision?

Class	**# Entities**
Product	368,386,801
Organization	176,365,220
Local Business	44,100,625
Place	31,748,612
Event	26,035,295
Job Posting	5,078,941
Movie	5,233,520
Hotel	2,537,089
College or University	2,392,797
Recipe	2,036,417
Music Album	1,594,941
Restaurant	1,480,443
Airport	1,145,323

Use Case 3: Training Data for Entity Matching / Link Discovery

- **Goal:** Find e-shops that offer a specific product in order to compare product prices
- **Observation:** Some e-shops annotates product IDs, many e-shops do not

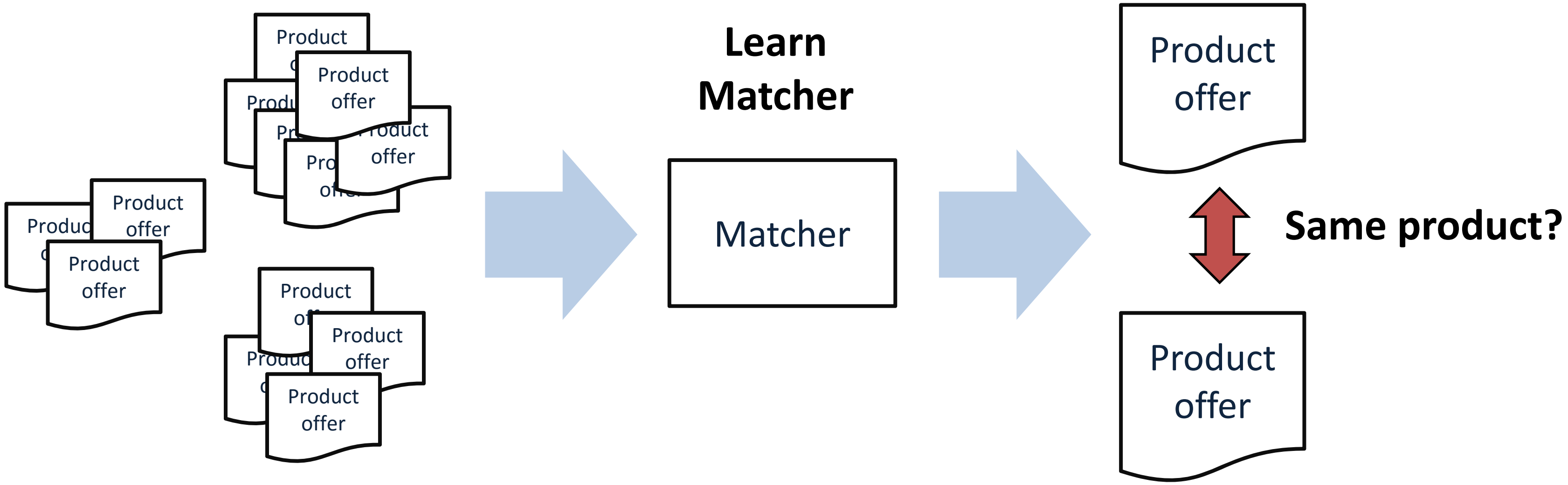
Properties	PLDs	
	#	%
schema:Product/name	754,812	92 %
schema:Product/description	520,307	64 %
schema:Product/sku	160,343	19 %
schema:Product/productID	47,088	5 %
schema:Product/mpn	12,882	1.6%
schema:Product/gtin13	7,994	1%

Learn How to Match Products using Schema.org Data as Supervision

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Clusters of offers having the same product ID

Unseen offers without product IDs



UNIVERSITY OF MANNHEIM
Data and Web Science Group

Data Cleansing for Building the Clusters

Filtering of product offers with annotated product identifiers	Removal of listing pages	Filtering by identifier value length	Cluster creation based on identifier value co-occurrence	Split wrong clusters due to category IDs
121M offers	**58M** offers	**26M** offers	**16.4M** clusters	**16.6M** clusters

All Languages
26.5M offers
79K distinct websites
16.6M clusters (products)

English Offers
16M offers
43K distinct websites
10M clusters (products)

Cluster Size Distribution by Category

Product Category	%Offers	%ID-Clusters	#ID-Clusters of Size 2	[3-4]	[5-10]	[11-20]	[>20]
Office	12.02%	10.90%	99,237	40,314	16,920	5,953	3,043
Jewelry	7.90%	7.79%	94,389	33,156	13,329	3,352	2,037
Clothing	7.52%	6.80%	95,006	49,085	30,384	3,285	1,866
Automotive	5.40%	5.90%	54,320	16,650	8,139	2,865	2,140
Beauty	5.50%	5.78%	70,984	27,636	10,568	2,115	1,070
Cell Phones & Acc.	5.74%	4.76%	48,659	15,870	5,162	1,085	878
Home & Kitchen	7.23%	4.68%	45,922	24,429	7,414	1,247	538
Luggage & Travel	4.06%	4.48%	36,211	14,401	6,399	1,198	957
Tools	3.67%	4.35%	28,236	12,033	4,407	1,248	1,042
CDs & Vinyl	3.88%	4.19%	42,261	17,666	6,013	1,417	663
Shoes	3.93%	4.11%	37,647	16,603	7,590	1,335	721
Camera & Photo	3.23%	3.47%	39,924	14,583	5,408	1,423	935
Grocery	3.39%	3.26%	39,372	17,109	5,889	2,154	716
Computers & Acc.	4.13%	3.20%	48,125	11,614	5,411	2,308	2,862
Digital Music	2.58%	3.03%	23,678	7,954	3,046	640	535
Other Electronics	2.43%	2.83%	31,034	11,649	4,412	977	427
Books	2.19%	2.81%	22,350	9,889	2,946	330	183
Video Games	2.43%	2.62%	18,918	8,256	3,419	938	779
Garden	1.82%	2.43%	13,360	4,898	1,764	475	366
Musical Instruments	2.02%	2.31%	22,263	5,182	1,684	550	486
Pet Supplies	1.83%	2.15%	13,589	7,605	2,974	620	440
Baby	1.49%	1.71%	16,853	5,509	1,894	458	254
Toys	0.95%	1.19%	10,525	3,120	1,016	258	189
Sports	0.86%	0.75%	8,216	3,460	1,234	314	372
Movies & TV	0.62%	0.71%	6,588	2,030	681	195	157
Health	0.65%	0.70%	7,261	6,857	1,165	189	113
Other Category	2.54%	3.07%	37,292	12,964	4,088	633	657
TOTAL	100%	100%	1,012,220	400,522	163,356	37,562	24,345

A. Primpeli, R. Peeters, C. Bizer: **The WDC Training Dataset and Gold Standard for Large-Scale Product Matching.** ECNLP 2019 Workshop @ WWW2019.

Training Subsets

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Category	#Positive Pairs	#Negative Pairs	s:name	s:description	specification table
Computers_small	20,444	21,676	100%	82%	55%
Cameras_small	7,539	9,093	100%	61%	4%
Watches_small	5,449	8,819	100%	48%	4%
Shoes_small	3,476	5,924	99%	36%	1%
Computers_large	**120,568**	**139,020**	100%	87%	60%
Cameras_large	**16,390**	**84,788**	100%	65%	5%
Watches_large	**7,729**	**65,516**	100%	55%	3%
Shoes_large	**3,476**	**36,224**	100%	35%	1%

Gold standard: 550 offer pairs per category

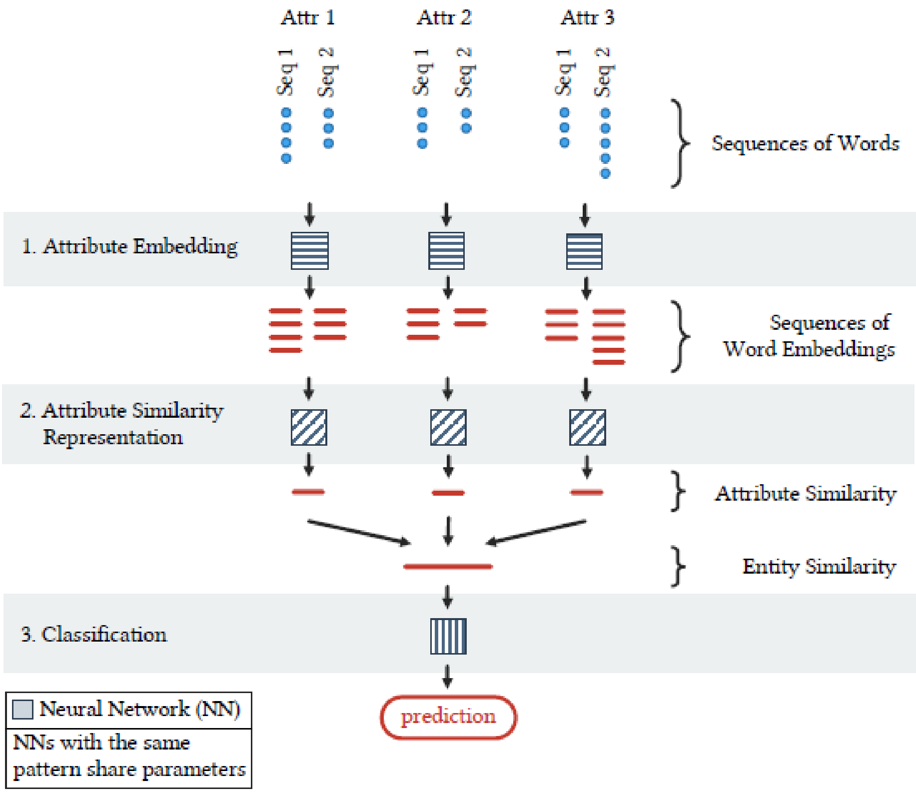
Results: Traditional Learning Methods

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Feature creation: word co-occurrence

Category	Classifier	Precision	Recall	F1
Computers_small	LinearSVC	0.75	0.94	0.84
Cameras_small	Random Forest	0.75	0.87	0.81
Watches_small	LinearSVC	0.74	0.91	0.81
Shoes_small	Log. Regression	0.72	0.95	0.82
Computers_large	Random Forest	0.92	0.95	**0.94**
Cameras_large	Dec. Tree	0.86	0.85	**0.86**
Watches_large	Random Forest	0.90	0.87	**0.88**
Shoes_large	Log. Regression	0.94	0.86	**0.90**

DeepMatcher



- word-based, character based
- pre-trained (word2vec, GloVe, fastText), learned

- attribute summarization: SIF, RNN, Attention, Hybrid
- attribute comparison: fixed distance, learnable distance

- classification: fully connected neural net

Mudgal, Li, et al.: **Deep Learning for Entity Matching: A Design Space Exploration.** SIGMOD 2018.

Results: DeepMatcher

Category	Model	Precision	Recall	F1
Computers_small	Combined SIF	0.91	0.98	0.95
Cameras_small	Combined RNN	0.95	0.93	0.94
Watches_small	Combined RNN	0.93	0.98	0.95
Shoes_small	Combined SIF	0.87	0.99	0.92
Computers_large	RNN	0.96	0.98	0.97
Cameras_large	RNN	0.94	0.98	0.96
Watches_large	RNN	0.96	0.99	0.97
Shoes_large	RNN	0.97	0.99	0.98

- Human-level matching performance just using web data for training.
- The 6.6% errors in the training data are averaged out in the learning.

Underexploited?

- If we can get human-level matching performance using just web data for training, why do product comparison portals still pay people to manually label offers for training?
- **Research questions**
  - Does this work for all product categories?
  - How good does it work for new products?

Use Case 3: Taxonomy Matching and Taxonomy Clustering

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- E-Shops annotate their product categorization

Relevant Properties	# PLDs	%
schema:Offer/category	62,170	7 % of all shops
schema:WebPage/BreadcrumbList	621,344	11% of all sites

- The marketing ideas behind these categorizations differ widely

Apparel > Summer > Mens > Jerseys

Philadelphia Eagles > Philadelphia Eagles Mens > Philadelphia Eagles Mens Jersey > over $60 s

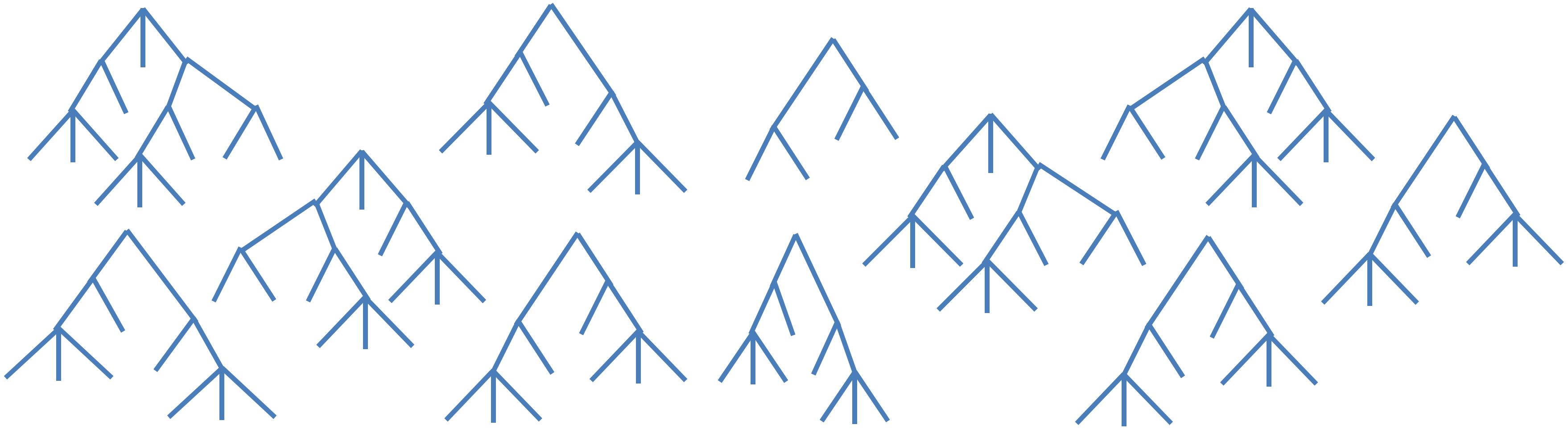
Home > Outdoor & Garden > Barbecues & Outdoor Living > Garden Furniture > Tables

Shop > Tables > Dining Tables > Aland Wood Tables

Cool Taxonomy Clustering Problem

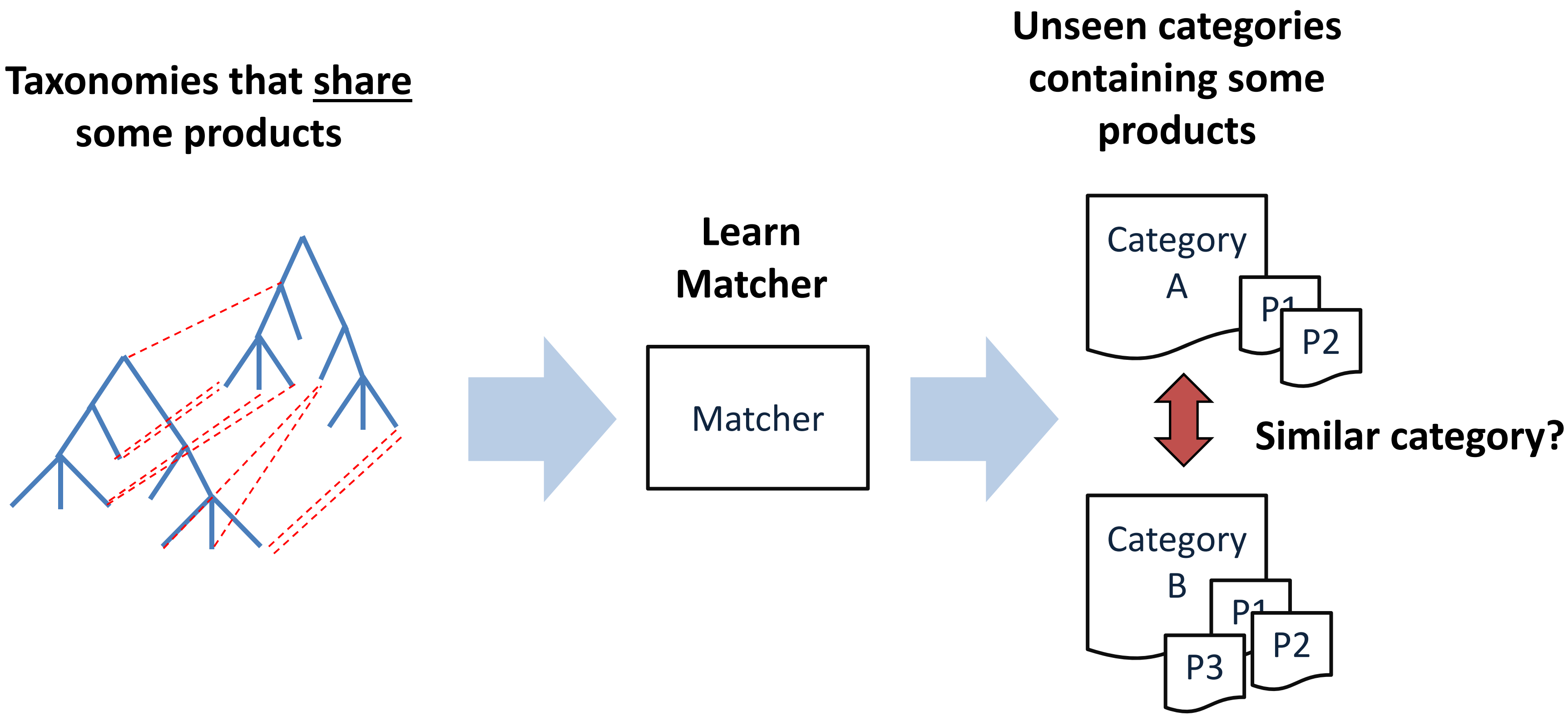
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- How to cluster **62,000 taxonomies** in order to discover the most frequent categorization ideas within a domain?
  - How do shops categorize action cameras?
  - What subcategories are frequently used for action cameras?



Meusel, Primpeli, Meilicke, Paulheim, Bizer**: Exploiting Microdata Annotations to Consistently Categorize Product Offers at Web Scale.** EC-Web 2015.

Learn How to Match Product Taxonomies using Schema.org Data as Supervision



Underexploited?

- Cool training data for the rather unexplored research challenge of large-scale taxonomy clustering
- **Research questions**
  - Which similarity metrics work best?
  - How to block 62,000 taxonomies in order to deal with the runtime complexity?
  - On which depths to start clustering in order to answer interesting business questions?

Use Case 4: Training Data for Sentiment Analysis

- 130,000 websites annotate reviews
- 2018 WDC corpus contains 13.5 million reviews which annotate schema:ratingValue and schema:reviewBody

```
<div  itemtype="http://schema.org/Product">
    <h1 itemprop="name">Signature Instant Dome 7</h1>
   Item # <span itemprop="productID">200734246807</span>
   <img itemprop="image" src="../20073424680.jpg">
   <span itemprop="review" itemtype="http://schema.org/Review">
        <span itemprop="ratingValue">5</span>
        <span itemprop="reviewBody">Great waterproof tent which
                                worked perfectly for me ….</span>
    </span>

</div>
```

Reviews by Type of Reviewed Item

schema.org Type	Reviews #	Reviews %	#Type entities
Product	6,371,735	46.96%	1,814,687
LocalBusiness	1,762,858	12.99%	455,200
Thing	1,405,100	10.35%	1,403,613
Organization	700,550	5.16%	233,216
Restaurant	465,208	3.43%	46,184
Hotel	368,831	2.72%	42,362
Book	350,826	2.59%	161,848
Service	217,817	1.61%	180,722
WebPage	144,441	1.06%	34,104
Place	60,348	0.44%	37,552
Article	59,943	0.44%	26,559
Event	59,916	0.44%	42,597
Person	57,740	0.43%	57,555
MobileApplication	41,610	0.31%	18,830
CreativeWork	4,072	0.03%	2,176

C. Bizer, A. Primpeli, R. Peeters: **Using the Semantic Web as a Source of Training Data**. Datenbank Spektrum, 2019.

Underexploited?

- Up to my knowledge, schema.org reviews from many websites are not used as training data for sentiment analysis yet.
- **Research questions**
  - For which cases does the training data help to improve sentiment analysis compared to other training corpora?
  - For which types of entities are there improvements?
  - For which languages?
  - Is the up-to-dateness of the training data relevant?

Conclusion: Semantic Annotations

- Comprehensive coverage of specific domains
  - many entities, many different websites
- Narrow set of attributes used
  - often requires additional parsing and specific cleansing
- Underexploited?
  - Information extraction: 1 paper, positive results
  - Entity matching: 1 paper, positive results
  - Taxonomy clustering: 0 experimental papers
  - Sentiment analysis: 0 experimental papers

2. Web Tables

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		
4	Cologne, North Rhine-Westphalia		
5	Frankfurt am Main, Hesse		
6			
7			
8			
9			
10			

NAME	WEBSITE
Aeroflot	www.aeroflot.ru/e
Air France	www.airfrance.fr
All Nippon Airways	www.ana.co.jp
Asiana Airlines	www.flyasiana.co
Cathay Pacific	www.cathaypacifi
China Airlines	www.china-airline

150 INTERNATIONAL AFFAIRS		2016	2017	2018	2019	2020
8.	End Funding for the United Nations Development Program (UNDP)	81	81	82	83	85
9.	End Funding for the U.N. Intergovernmental Panel on Climate Change (IPCC)	10	10	10	10	11
10.	Eliminate the U.S. Trade and Development Agency (USTDA)	56	56	56	57	58
11.	Reform Food Aid Programs	168	168	168	170	173
12.	Eliminate Export-Import Bank					
13.	Eliminate the Overseas Private Investment Corporation (OPI					
14.	Eliminate Funding for the Uni Nations Population Fund (UN					
SUBTOTALS						

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

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	Cupid	Cupid Shuffle
6	Bon Jovi	Livin' On A Prayer
7	Beyonce	Single Ladies (Put A Ring On It)
8	Diamond, Neil	Sweet Caroline (Good Times Never Seemed So Good)
9	Morrison, Van	Brown Eyed Girl
10	Def Leppard	Pour Some Sugar On Me
11	B-52's	Love Shack
12	Lmfao Feat. Lauren Bennett And Goon Rock	Party Rock Anthem
13	Jackson, Michael	Billie Jean
14	DJ Casper	Cha Cha Slide

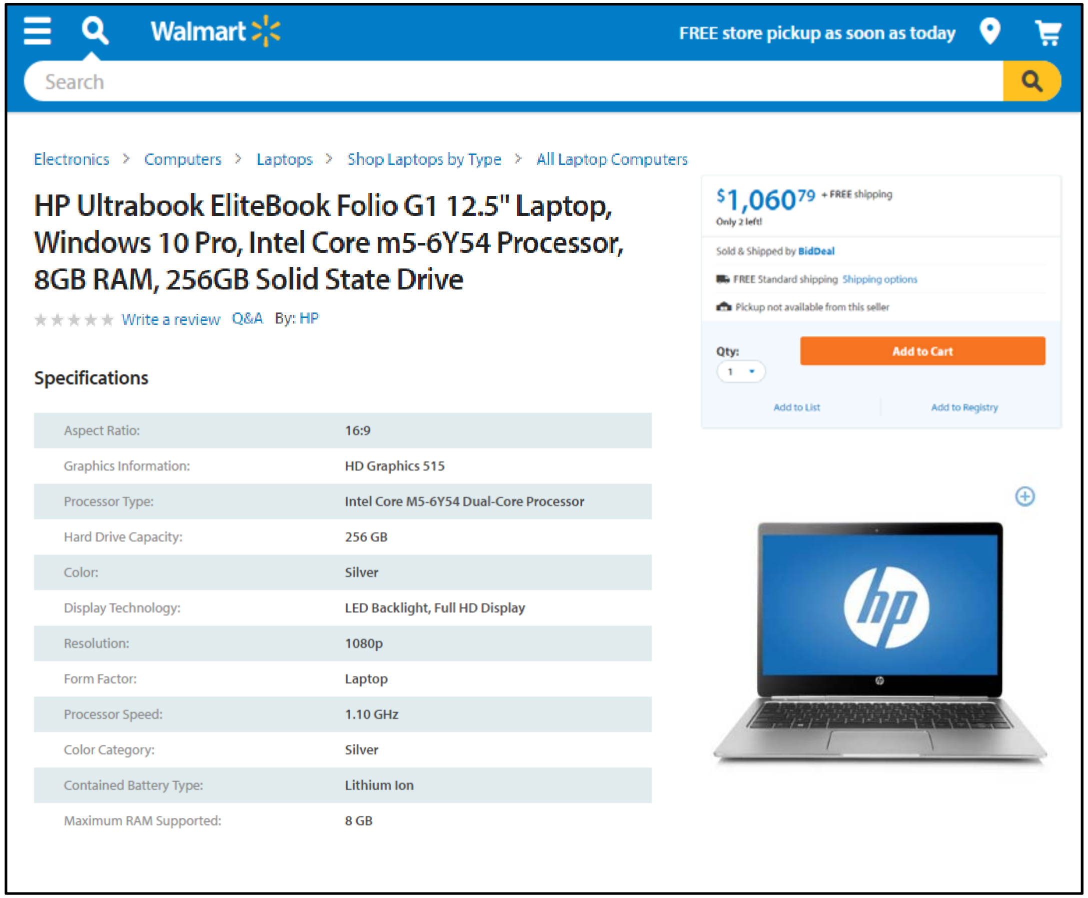
https://research.google.com/tables

Attribute/Value Table together with schema.org Annotations

s:breadcrumb

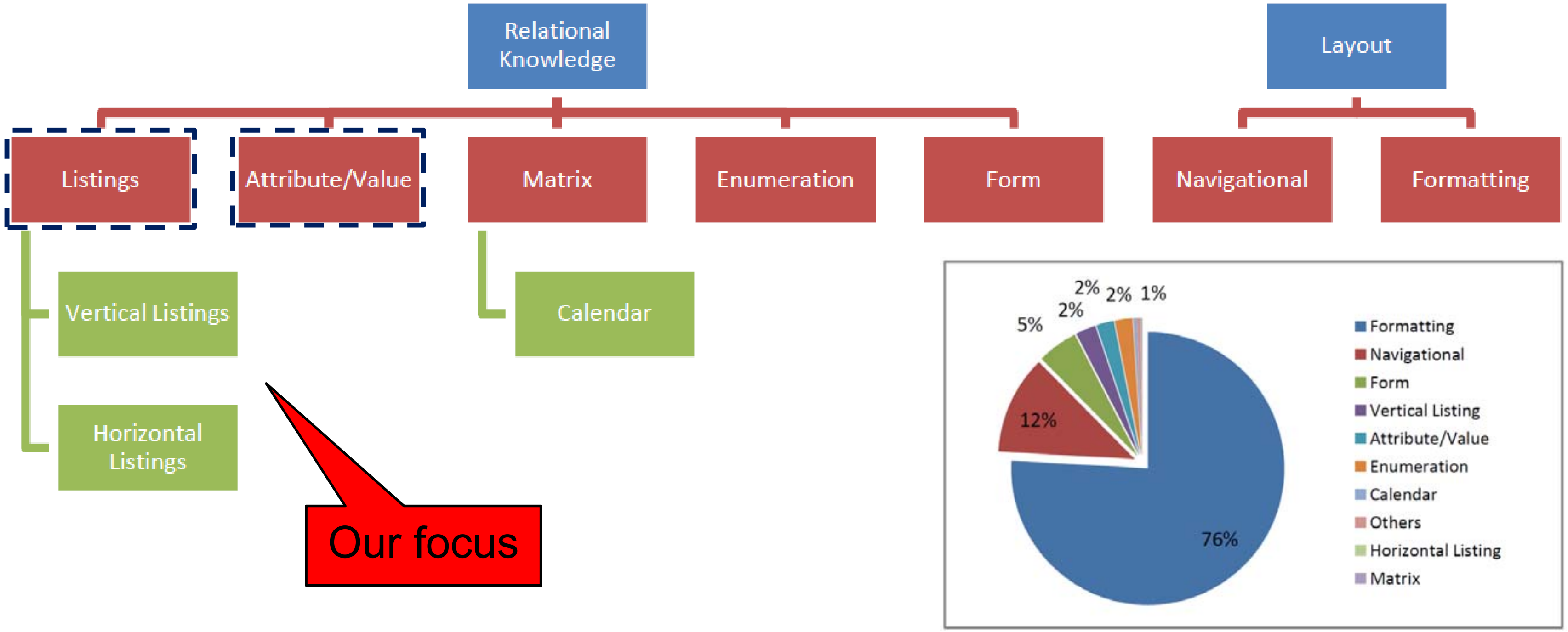
s:name

Attribute/Value HTML Table



Qui, et al.: **DEXTER: Large-Scale Discovery and Extraction of Product Specifications on the Web.** VLDB 2015.
Petrovski, et al: **The WDC Gold Standards for Product Feature Extraction and Product Matching.** ECWeb 2016.

Web Tables



In corpus of 14B raw tables, 154M are relational tables (1.1%).

Cafarella et al. 2008

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

Web Data Commons – Web Tables Corpus

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Large public corpus of Web tables
- extracted from Common Crawl 2015 (1.78 billion pages)
  - 90 million relational tables (0.9%)
  - 139 million attribute/value tables (1,3%)
  - selected out of 10.2 B raw tables
  - 1TB zipped
- http://webdatacommons.org/webtables/

Lehmberg, Ritze, Meusel, Bizer: **A Large Public Corpus of Web Tables containing Time and Context Metadata.** WWW2016 Companion.

Topical Profiling: Websites contributing many tables

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

Usage Scenario:
Knowledge Base Completion

Class in KB

	A_1	A_2	A_3	A_4	…	A_N	?	?	?	…
E_1	?	?	?				?	?	?	?
E_2		?			?		?	?	?	?
E_3							?	?	?	?
…			?				?	?	?	?
E_m	?		?	?	?		?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
…	?	?	?	?	?	?	?	?	?	?

Slot Filling:
add facts for existing instances and existing properties

Schema Expansion:
add and fill new attributes

Entity Expansion:
- add new instances
- and their descriptions (values for existing attributes)

Use Case 1: Slot Filling

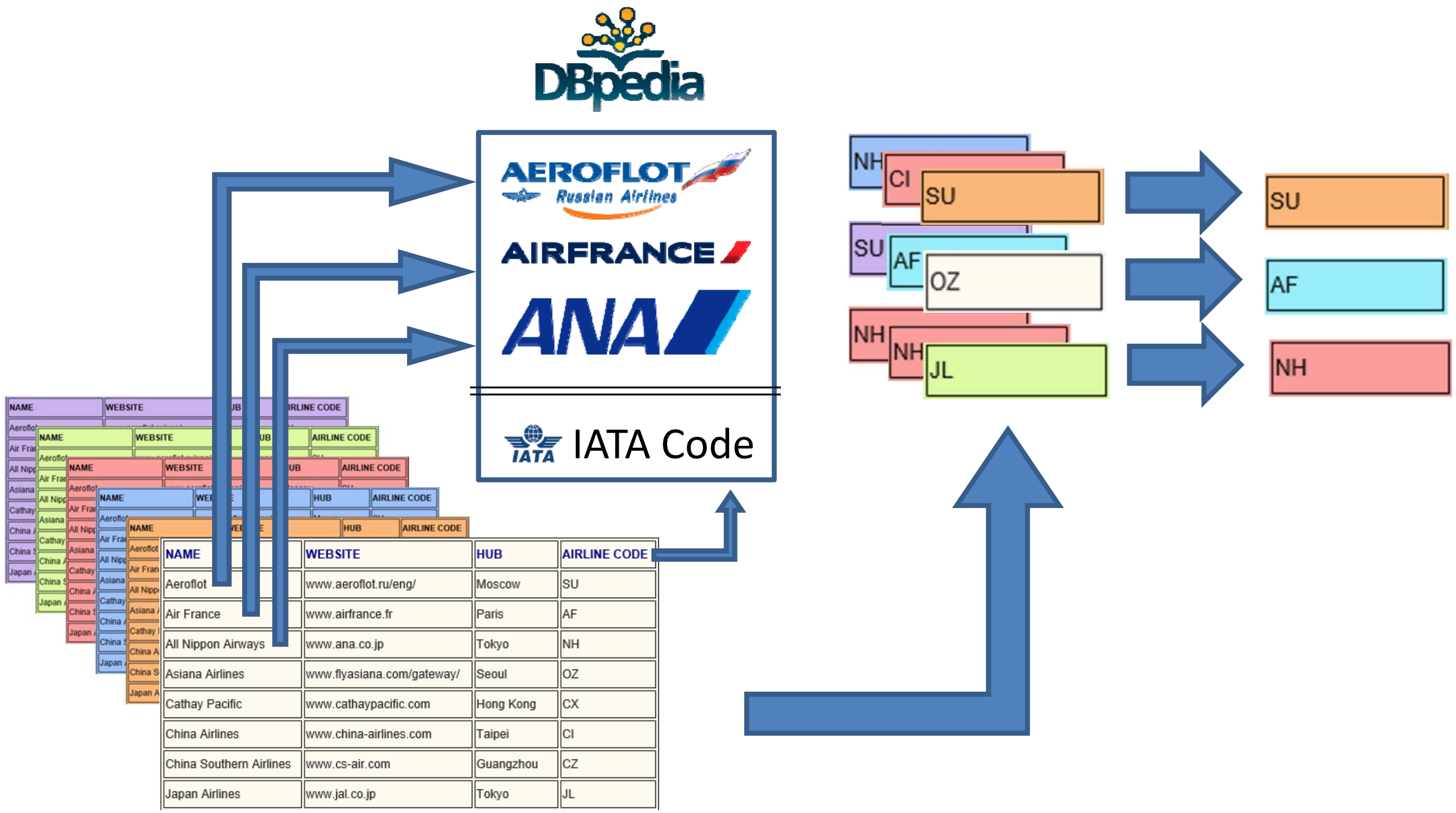
Class in KB

	A_1	A_2	A_3	A_4	…	A_N
E_1	?	?	?			
E_2		?			?	
E_3						
…			?			
E_m	?		?	?	?	

Slot Filling: Two Steps

1. Matching

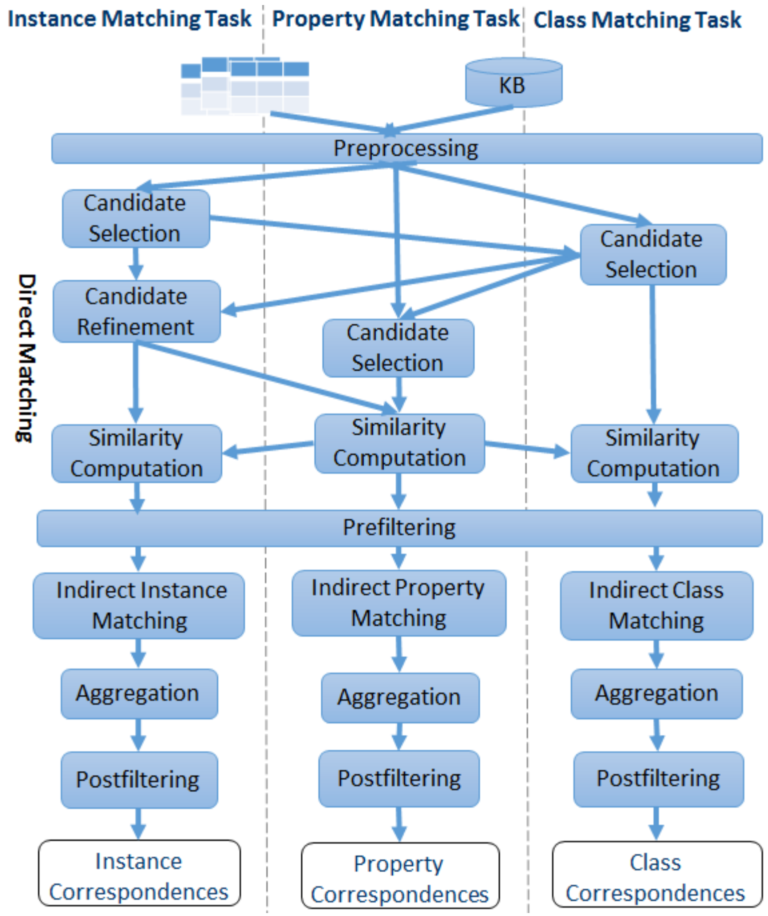
2. Data Fusion



Holistic Entity- and Schema Matching

- T2K+ matches
  - millions of web tables to the
  - DBpedia knowledge base
- Exploits
  - wide range of features
  - synergies between matching tasks
  - table corpus for indirect matching
  - patterns from KB for post filtering
- Results

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



Dominique Ritze, Christian Bizer: **Matching Web Tables To DBpedia – A Feature Utility Study.** EDBT 2017.

Matching Small Web Tables: Table Stitching

- Tables with the same schema appear many times on the same website
  - as large original tables are split for human consumption
  - 97% of all websites use between 1 and 10 different schemata [VLDB'17]
  - Stitching merges all web tables on one website having the same schema

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
California	4,950	0.31	1.64	$60.39	$125,620
Virginia	2,550	0.68	3.59	$60.96	$126,800
Maryland	2,490	0.94	4.99	$54.38	$113,110
Texas	1,810	0.15	0.82	$47.52	$98,830
New Jersey	1,530	0.39	2.04	$60.52	$125,880

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
New Mexico	890	1.10	5.83	$54.79	$113,950
Maryland	2,490	0.94	4.99	$54.38	$113,110
Rhode Island	430	0.91	4.79	$52.31	$108,810
District of Columbia	640	0.91	4.80	$59.37	$123,490
Virginia	2,550	0.68	3.59	$60.96	$126,800

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
Oregon	250	0.14	0.74	$74.14	$154,220
New York	920	0.10	0.53	$70.26	$146,150
Washington	(8)	(8)	(8)	$64.40	$133,950
New Hampshire	170	0.26	1.40	$61.67	$128,280
Virginia	2,550	0.68	3.59	$60.96	$126,800

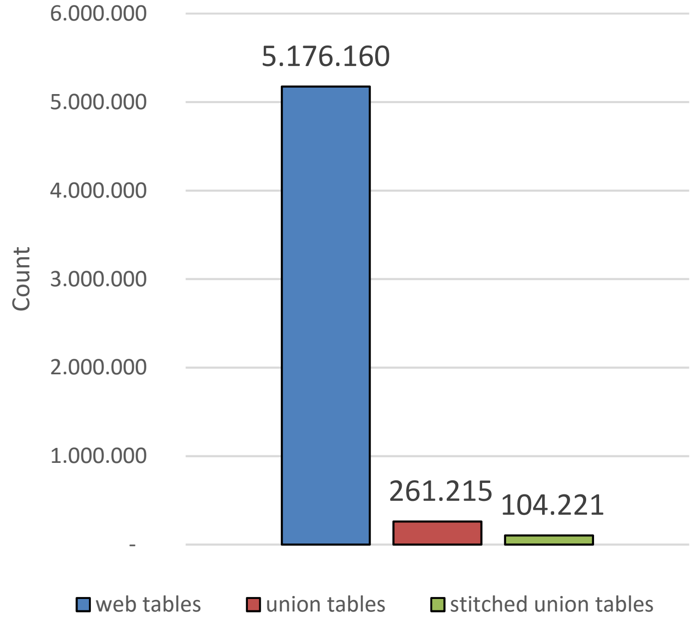
$$S = \bigcup_{T_i \in \mathcal{T}} T_i$$

State	Employment (1)	Employment per thousand jobs	Location quotient (9)	Hourly mean wage	Annual mean wage (2)
California	4,950	0.31	1.64	$60.39	$125,620
Virginia	2,550	0.68	3.59	$60.96	$126,800
Maryland	2,490	0.94	4.99	$54.38	$113,110
Texas	1,810	0.15	0.82	$47.52	$98,830
New Jersey	1,530	0.39	2.04	$60.52	$125,880
New Mexico	890	1.10	5.83	$54.79	$113,950
Maryland	2,490	0.94	4.99	$54.38	$113,110
Rhode Island	430	0.91	4.79	$52.31	$108,810
District of Columbia	640	0.91	4.80	$59.37	$123,490
Virginia	2,550	0.68	3.59	$60.96	$126,800
Oregon	250	0.14	0.74	$74.14	$154,220
New York	920	0.10	0.53	$70.26	$146,150
Washington	(8)	(8)	(8)	$64.40	$133,950
New Hampshire	170	0.26	1.40	$61.67	$128,280
Virginia	2,550	0.68	3.59	$60.96	$126,800

Oliver Lehmberg, Christian Bizer: **Stitching web tables for improving matching quality**. *VLDB Endowment*, 2017.

Effect of Table Stitching

- Strongly reduces the amount of tables
  - 19.8 web tables on average per union table
  - 46.9 web tables on average per stitched union table
- Attribute matching F1:
  - original web tables 0.35
  - union tables 0.53
  - stitched union tables **0.67**
  - expected from T2K+ 0.70



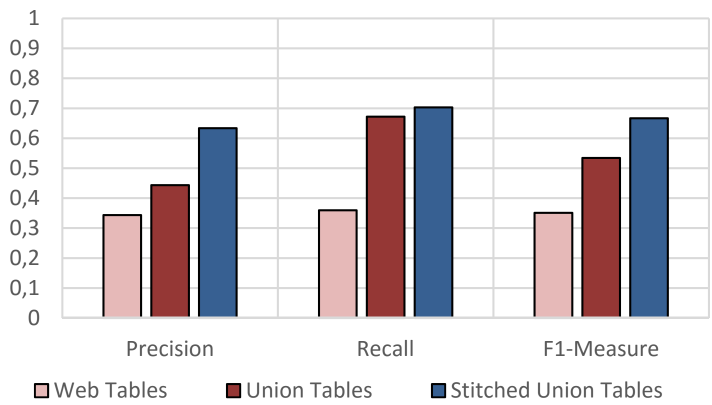


Table to DBpedia Matching Results

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- about 1 million out of 30 million tables match DBpedia (~3%)
  - 13,726,582 row-to-instance correspondences (~0.7%)
  - 562,445 attribute-to-property correspondences
  - 301,450 tables with attribute correspondences (ca. 32%)

    ➔ 8 million triples
- Content variety
  - 274 different classes, 721 unique properties
  - 717,174 unique instances (15.6% of DBpedia)

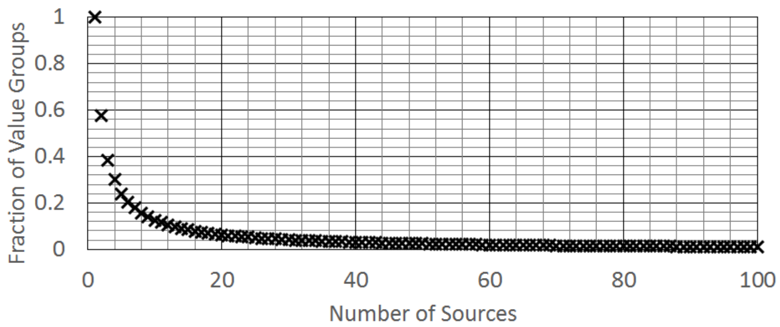
Ritze, Lehmberg, Oulabi and Bizer: **Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases.** WWW2016.

Data Fusion Results

- Fusion Results

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

- Group Sizes
  - Large groups (Head): values likely already exist in the KB
  - Small groups (Tail): often new value but difficult to fuse correctly



Fused Triples from Web Tables by DBpedia Class

DBpedia Class	Overlap with existing values	New values
Person	**117 522**	**15 050**
Athlete	84 562	9 067
Artist	2 019	427
Office Holder	3 465	510
Politician	3 124	1 167
Organisation	**20 522**	**7 903**
Company	6 376	2 547
Sports Team	790	132
Educational Inst.	8 844	3 132
Broadcaster	4 004	1 924
Work	**189 131**	**27 867**
Musical Work	118 511	8 427
Film	29 903	12 143
Software	17 554	2 766
Place	**32 855**	**9 871**
Populated Place	16 604	6 704
Country	2 084	433
Architectural Struct.	10 441	1 775
Species	9 016	1 429

What is going wrong?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

UNITED STATES DEPARTMENT OF LABOR | A to Z Index | FAQs | About BLS | Contact Us | Subscribe to E-mail Updates | GO

BUREAU OF LABOR STATISTICS | Follow Us | What's New | Release Calendar | Blog | Search BLS.gov

Home | Subjects | Data Tools | Publications | Economic Releases | Students | Beta

Occupational Employment Statistics

OES | SHARE ON: | PRINT:

BROWSE OES
- OES HOME
- OES OVERVIEW
- OES NEWS RELEASES
- OES DATA
- OES CHARTS
- OES MAPS
- OES PUBLICATIONS

May 2017 State Occupational Employment and Wage Estimates

California

Occupation code	Occupation title (click on the occupation title to view its profile)	Level	Employment	Employment RSE	Employment per 1,000 jobs	Location quotient	Median hourly wage	Mean hourly wage	Annual mean wage
11-3021	Computer and Information Systems Managers	detail	53,270	2.2%	3.191	1.24	$77.80	$84.03	$174,790
15-0000	**Computer and Mathematical Occupations**	major	590,550	1.2%	35.373	1.18	$48.31	$50.66	$105,380
15-1111	Computer and Information Research Scientists	detail	5,750	9.0%	0.345	1.76	$59.20	$61.80	$128,530
15-1121	Computer Systems Analysts	detail	72,980	3.2%	4.371	1.07	$47.08	$49.45	$102,860
15-1131	Computer Programmers	detail	34,050	3.7%	2.040	1.17	$45.01	$46.28	$96,270

https://www.bls.gov/oes/2017/may/oes_ca.htm

What is going wrong?

- Matching and fusion methods expect binary relations
- But many relations in Web tables are not binary!

State	Avg. Sal.
California	$128,805
New York City	$113,689
Alabama	

State	Avg. Sal.
California	$63,805
New York City	$53,472
Alabama	$39,180


UNITED STATES DEPARTMENT OF LABOR

BUREAU OF LABOR STATISTICS

Occupational Employment Statistics

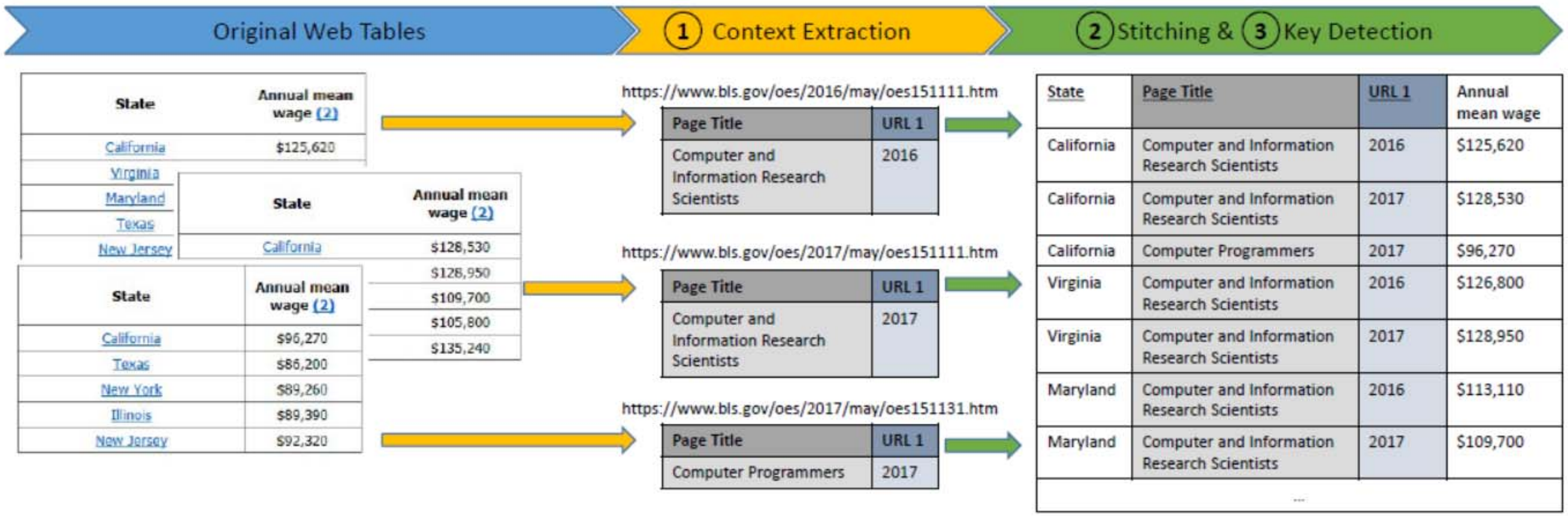
May 2017 State Occupational Employment and Wage Estimates

California

Occupation code	Occupation title (click on the occupation title to view its profile)	Level	Employment	Employment RSE	Employment per 1,000 jobs	Location quotient	Median hourly wage	Mean hourly wage	Annual mean wage
11-3021	Computer and Information Systems Managers	detail	53,270	2.2%	3.191	1.24	$77.80	$84.03	$174,790
15-0000	Computer and Mathematical Occupations	major	590,550	1.2%	35.373	1.18	$48.31	$50.66	$105,380
15-1111	Computer and Information Research Scientists	detail	5,750	9.0%	0.345	1.76	$59.20	$61.80	$128,530
15-1121	Computer Systems Analysts	detail	72,980	3.2%	4.371	1.07	$47.08	$49.45	$102,860
15-1131	Computer Programmers	detail	34,050	3.7%	2.040	1.17	$45.01	$46.28	$96,270

Synthesizing N-ary Relations from Web Tables

– Exploiting page context: Title and URL
– Stitching tables from the same site

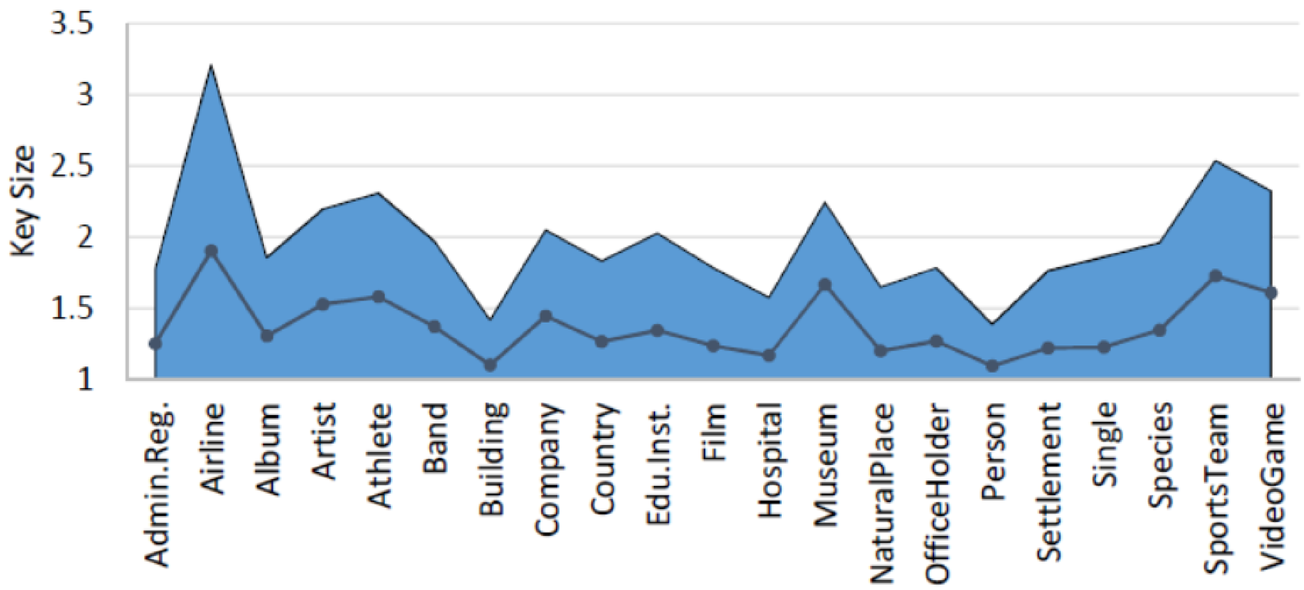


Oliver Lehmberg, Christian Bizer: **Synthesizing N-ary Relations from Web Tables.** WIMS2019.

Amount of N-ary Relations

Results for the WDC2015 corpus:

– 37.50% of all relations are non-binary
– 50.47% of these require context attributes



Time dimension is often part of keys but also many other attributes

Oliver Lehmberg, Christian Bizer: **Profiling the Semantics of N-ary Web Table Data.** SBD Workshop @ SIGMOD 2019.

Use Case 2: Adding Long-Tail Entities to Knowledge Base

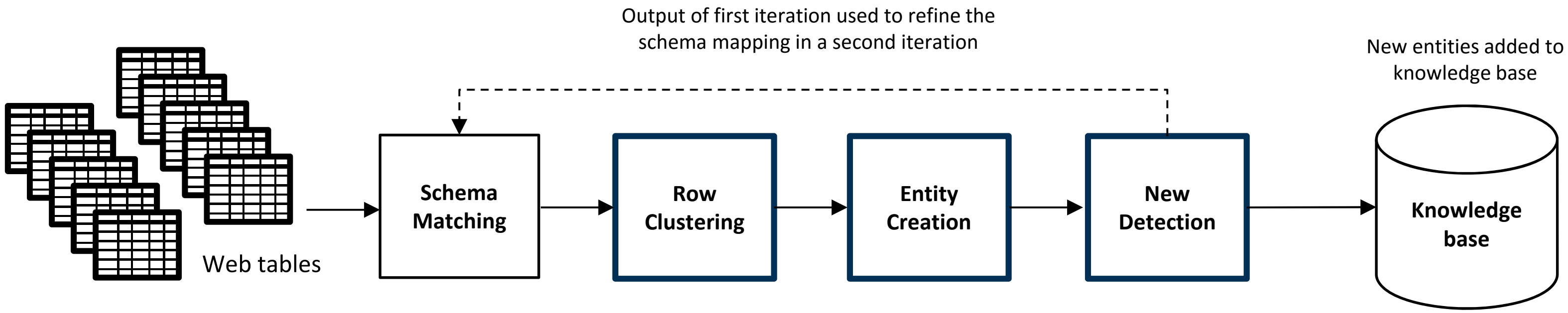
Class in KB

	A_1	A_2	A_3	A_4	…	A_N	?	?	?	…
E_1	?	?	?				?	?	?	?
E_2		?			?		?	?	?	?
E_3							?	?	?	?
…			?				?	?	?	?
E_m	?		?	?	?		?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?
…	?	?	?	?	?	?	?	?	?	?

Entity Expansion:

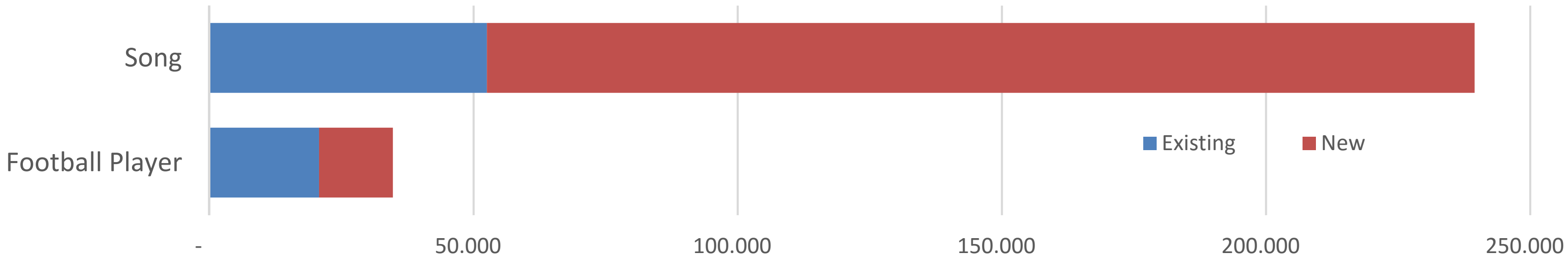
1. Find new entities
2. Compile descriptions of these new entities

Long-Tail Entity Expansion Pipeline



Yaser Oulabi, Christian Bizer: **Extending cross-domain knowledge bases with long tail entities using web table data.** EDBT 2019.

Entity Expansion Results



Class	New Entities	New Facts	New Entities Accuracy	New Facts Accuracy
Football Player	13,983 (+67%)	43,800 (+32%)	0.60	0.95
Song	186,943 (+356%)	393,711 (+125%)	0.70	0.85

– 43 % of errors caused by wrong attribute-to-property matching
– 35 % of errors are a result of outdated or wrong data in tables

Conclusion: Web Tables

- Extremely wide set of schemata with very specific attributes
  - many n-ary relations with complex keys / time dimension
- Table context is often required to understand the data
  - page title, URL, headings, text?
- Everything works OK, but not good enough yet
  - fact precision slot filling: 0.64
  - new entities accuracy: 0.60-0.70
- For researchers: Rewarding but challenging domain
- For applications: Human in the loop still required for verification

Conclusion

Dimension	Semantic Annotations	Web Tables
Classes	Few classes, deep coverage (Products, Local Business, …)	Wide range of topics (Sports, Statistics, …)
Attributes	Small set of rather generic attributes	Wide range of very specific attributes
Challenges	Parsing of below 1NF attribute values	Understanding the semantics of attributes
Opportunities	Resource creation **Large-scale training**	Resource creation
Underexploited	Yes	Yes, once challenges are handled