UNIVERSITY OF MANNHEIM

Team Project HWS 2022

Table Annotation using Deep Learning

Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web Data Integration
  - Data and Web Mining
  - Deployment of Data Web Technologies
- Room: B6 - B1.15
- eMail: christian.bizer@uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30

Hallo

- **M. Sc. Wi-Inf. Keti Korini**
- Graduate Research Associate
- Research Interests:
  - Table Annotation using Deep Learning
  - Schema Matching
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de

Agenda of Today‘s Kickoff Meeting

1. A round of introductions: You and Your Experience
2. Introduction and Project Goals
3. The WDC Schema.org Table Annotation Benchmark
4. Organization
5. Specific Subtasks
6. Schedule
7. Formal Requirements

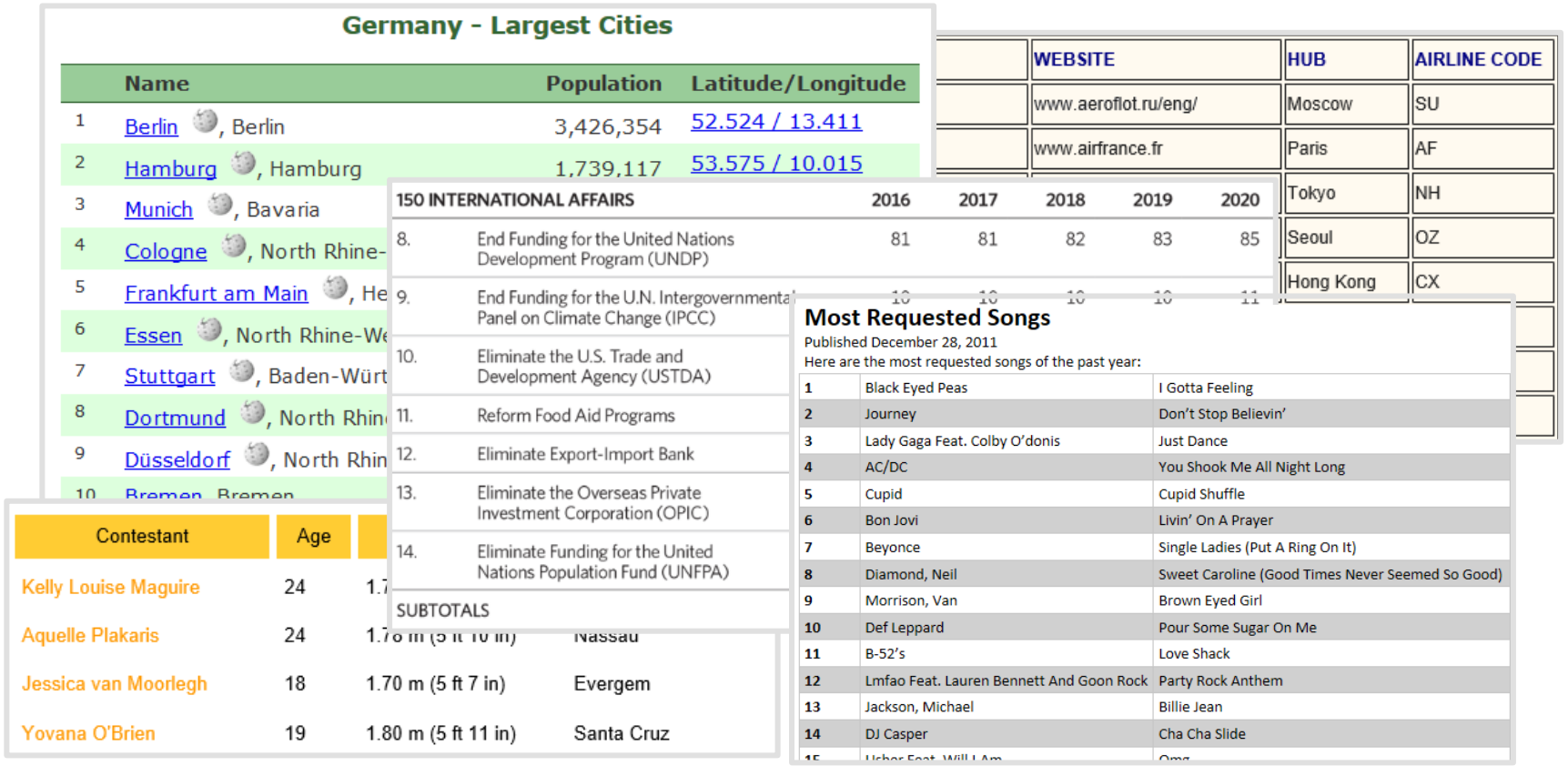
You and Your Experience

- A Short Round of Introductions
  - What are you studying? Which semester?
  - Which DWS courses did you already attend?
  - What are your programming and data wrangling skills?
  - Did you already work on any data integration or cleansing projects?
- Participants

1. Poudel, Subash
2. Der, Reng
3. Chen, Chun-Yi
4. Hsieh, I-Chen
5. Abobaker, Munir

2. Introduction: The Web as Source of Structured Data

There are hundreds of millions of high-quality tables available on the Web and in Wikipedia.



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-		
5	Frankfurt am Main, He		
6	Essen, North Rhine-We		
7	Stuttgart, Baden-Württ		
8	Dortmund, North Rhin		
9	Düsseldorf, North Rhin		
10	Bremen, Bremen		

	WEBSITE	HUB	AIRLINE CODE
	www.aeroflot.ru/eng/	Moscow	SU
	www.airfrance.fr	Paris	AF
		Tokyo	NH
		Seoul	OZ
		Hong Kong	CX

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)					
11.	Reform Food Aid Programs					
12.	Eliminate Export-Import Bank					
13.	Eliminate the Overseas Private Investment Corporation (OPIC)					
14.	Eliminate Funding for the United Nations Population Fund (UNFPA)					
	SUBTOTALS					

Contestant	Age		
Kelly Louise Maguire	24	1.7	
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

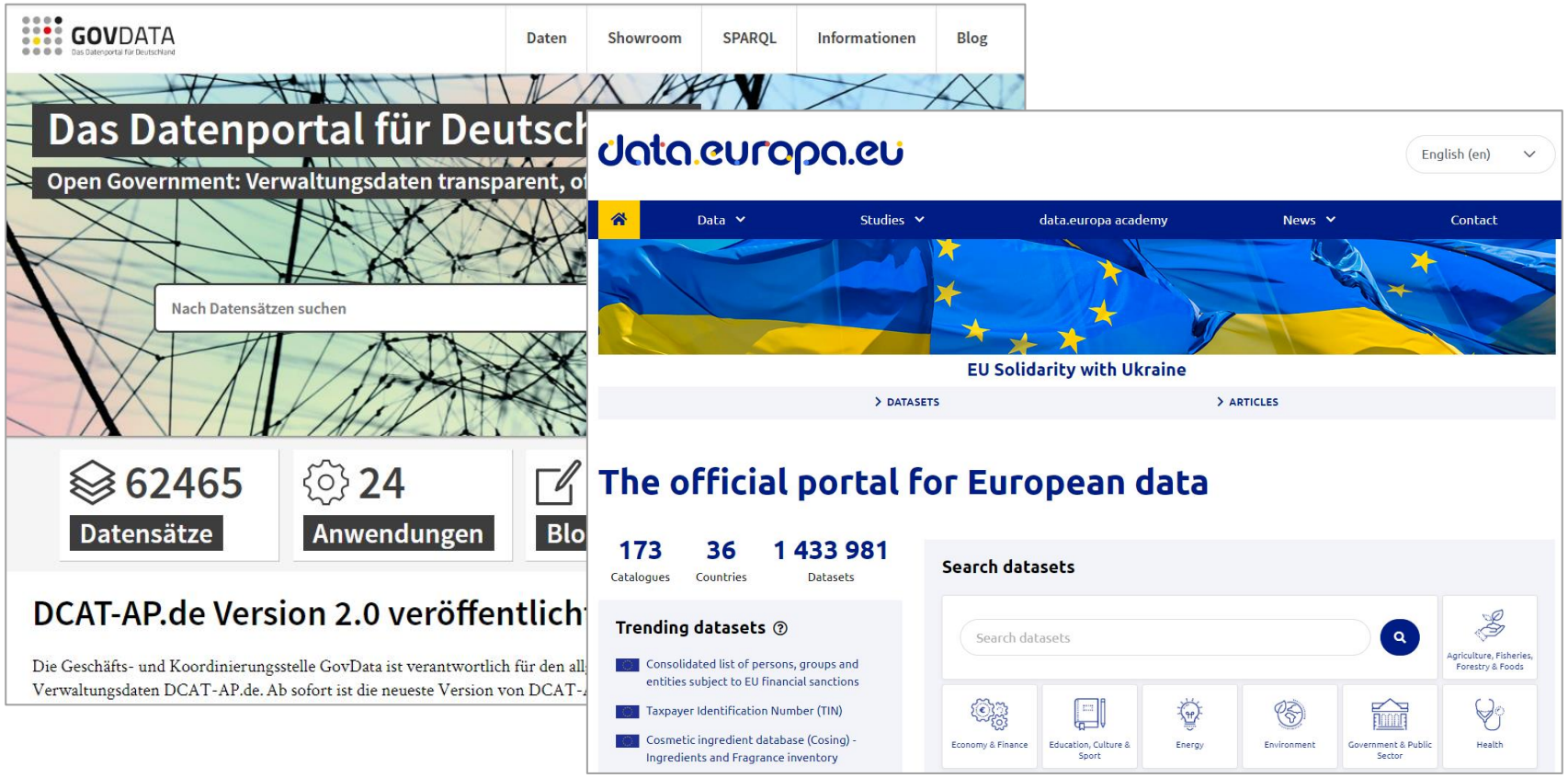
Published December 28, 2011

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

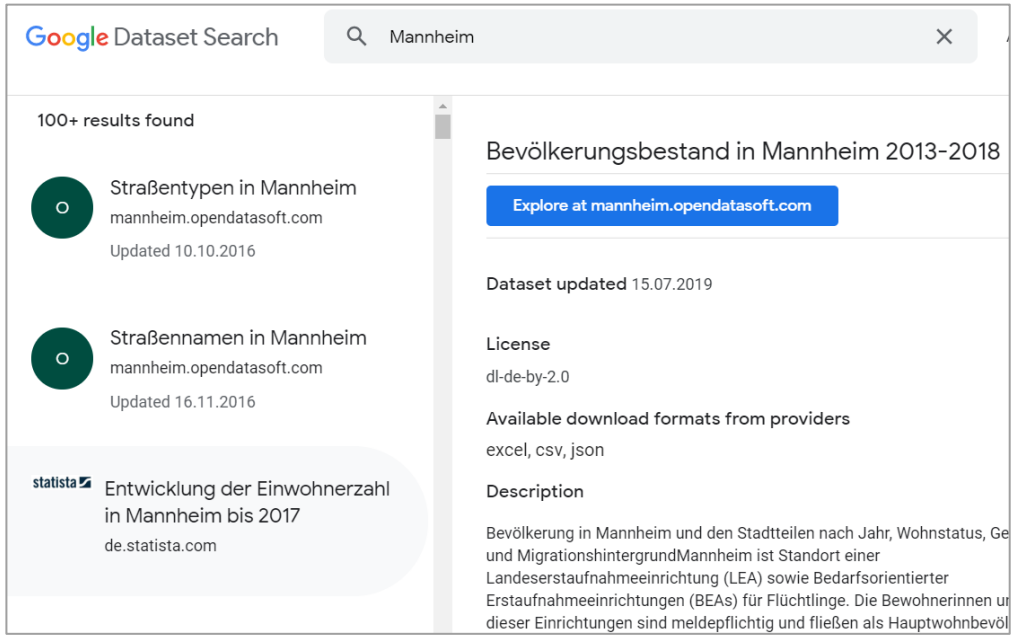
The Web as Source of Structured Data

In addition, there are millions of tables available via public data portals.



Challenge for Data Search

For providing advanced dataset search, one needs to understand the schemata of the tables

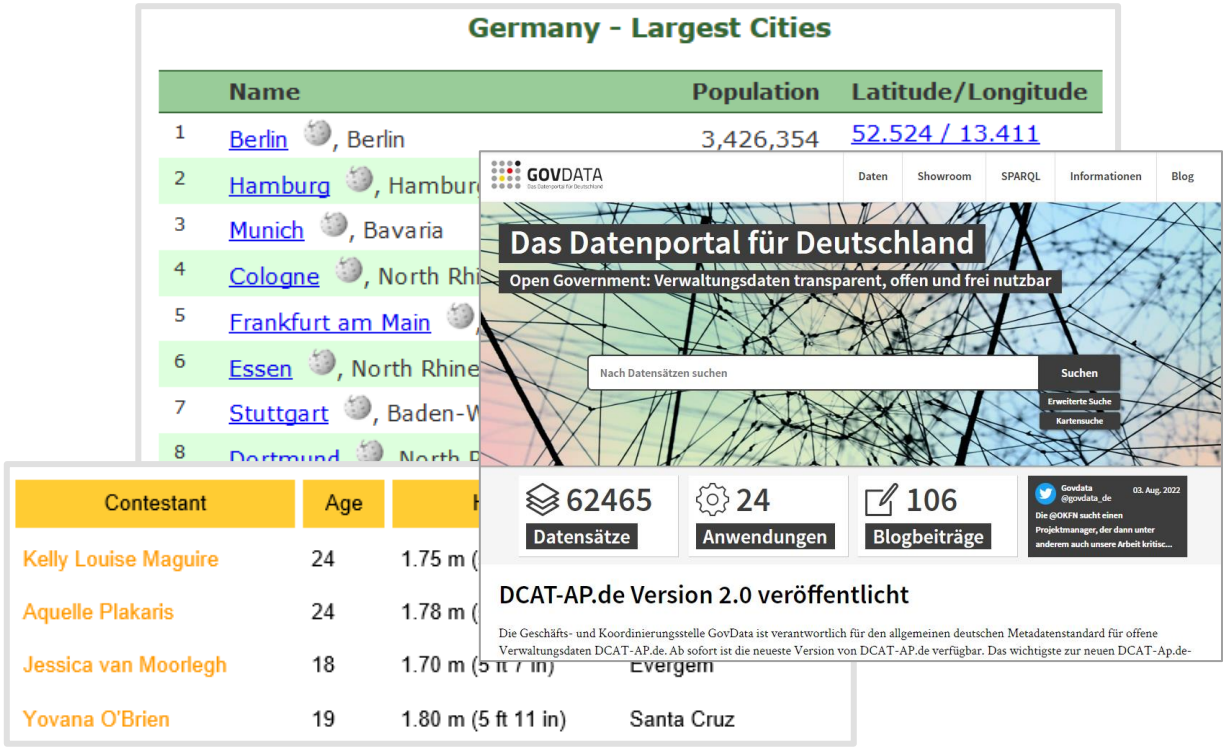


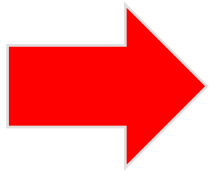
Give me datasets describing movies that provide directors and release years?

Give me all datasets containing population numbers for German cities?

Challenge for Data Integration

For using data from these tables to complete databases or knowledge graphs, one needs to understand the schema of the tables.





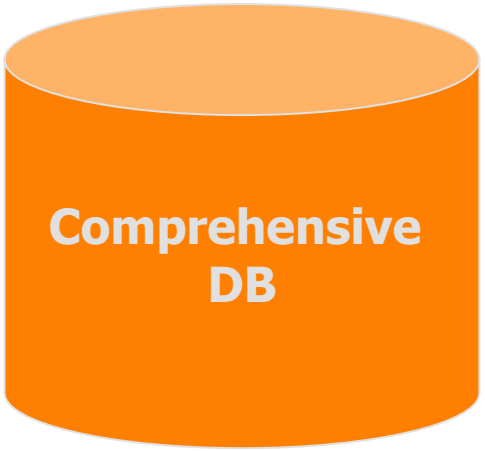


Table Annotation

Goal: Annotate tables in a large table corpus with concepts from a knowledge graph or shared vocabulary

- Enabling step for data search and data integration
- Challenges: Heterogeneous or unknown headers, heterogeneous content

film	director	producer	country
???	???	???	???
Happy Feet	George Miller, Warren Coleman, Judy Morris	Bill Miller, George Miller, Doug Mitchell	USA
Cars	John Lasseter, Joe Ranft	Darla K. Anderson	UK
Flushed Away	David Bowers, Sam Fell	Dick Clement, Ian La Frenais, Simon Nye	France

annotate

Table Annotation Tasks

1. cell entity annotation (CEA)
2. column type annotation (CTA)
3. columns property annotation (CPA)
4. row annotation
5. table type detection

film	director	producer	country
???	???	???	???
Happy Feet	George Miller, Warren Coleman, Judy Morris	Bill Miller, George Miller, Doug Mitchell	USA
Cars	John Lasseter, Joe Ranft	Darla K. Anderson	UK
Flushed Away	David Bowers, Sam Fell	Dick Clement, Ian La Frenais, Simon Nye	France

Papers with Code: Table Annotation: https://paperswithcode.com/task/table-annotation

Task 1: Column Type Annotation

Column Type Annotation (CTA): Annotation of table columns with the type of the entities contained in the column.

Entity types capture more domain semantics than **data types:**

- **Data types:** string, integer, Boolean …
- **Entity types:** country, product, person, duration, distance, weight, …

→ Usually approached as a multi-class classification problem

- **Input:** column + table context
- **Output:** Column type label from set of possible labels

Schema.org types and properties

Hotel/name	streetAddress	addressLocality	Country	currency
Lau's Gateway	209 Main Street	Alofi	NU	NZD
Radisson Blu Hotel, Nice	223 Promenade Des Anglais	Nice	FR	EUR
Phoenix Park Hotel	38-39 Parkgate Street Dublin 8	Dublin	IE	EUR

Task 2: Columns Property Annotation

Columns Property Annotation (CPA): Annotation of the relationship between the subject column and a second column.

Subject/label column contains the name of the entity described in a row:

- often the left-most column in the table,
- the other columns describe attributes of the named entity

→ Treated as multi-class classification problem

- **Input:** column pair + table context
- **Output:** columns property label

Subject column

film	director	producer	country
???	???	???	???
Happy Feet	George Miller, Warren Coleman, Judy Morris	Bill Miller, George Miller, Doug Mitchell	USA
Cars	John Lasseter, Joe Ranft	Darla K. Anderson	UK
Flushed Away	David Bowers, Sam Fell	Dick Clement, Ian La Frenais, Simon Nye	France

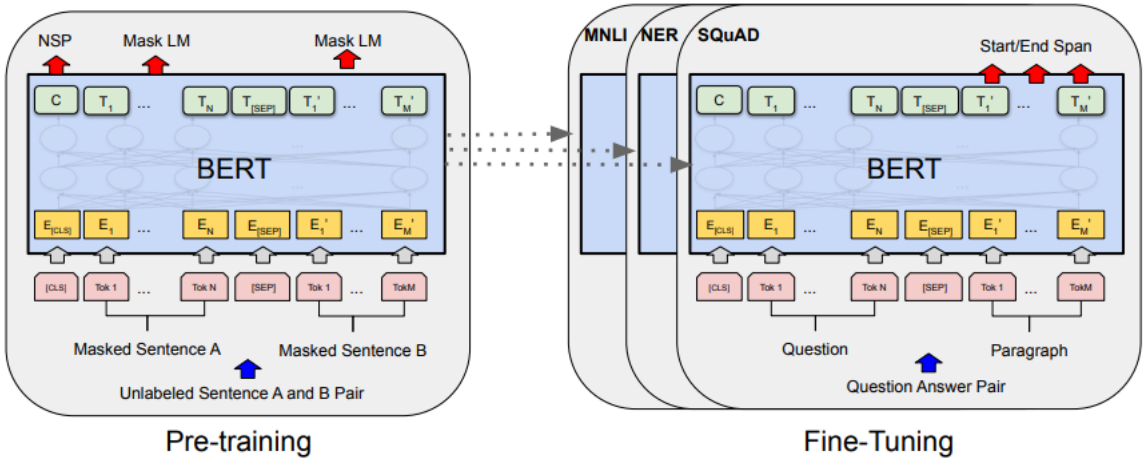
Transformer in NLP

Transformer architectures like BERT had large impact on NLP

- stacked encoder layers with self-attention mechanism
- contextual representation of tokens within a sequence of tokens
- [CLS] token that summarizes the sequence and can be used for classification

Pre-training / Fine-tuning paradigm

- pre-training: self-supervised masked language modeling on large text-corpora
- fine-tuning: further training on task-specific data

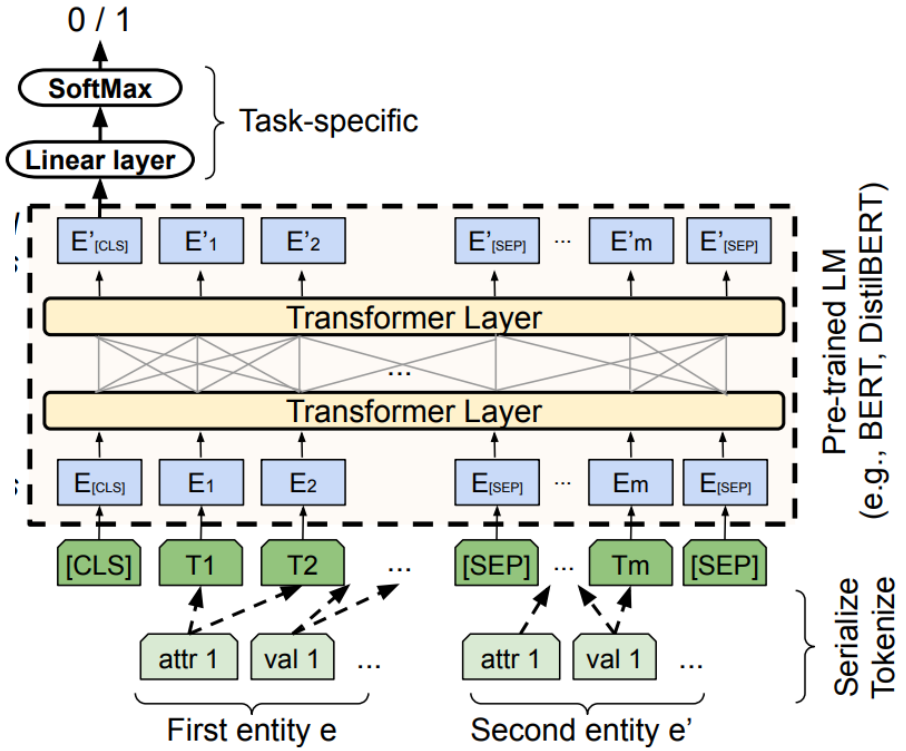


→ shown to work extremely well for a variety of problems

→ growing body of work regarding data integration using Transformers (see references)

Transformers for Entity Matching: DITTO (2021)

- applies BERT, DistilBERT, RoBERTa for entity matching
- adds methods for entity summarization, training data augmentation
- Entity serialization for BERT
  - Pair of entity descriptions are turned into single sequence
  - [CLS] Entity Description 1 [SEP] Entity Description 2 [SEP]
  - Entity Description = [COL] $attr_1$ [VAL] val_1 . . . [COL] $attr_k$ [VAL] val_k
- [CLS] token summarizes the pair of entities
- linear layer on top of [CLS] token for matching decision
- uses augmentation to increase amount of training data



Yuliang, et al: **Deep entity matching with pre-trained language models.** PVLDB 2021.

DITTO: Evaluation

Type	Dataset	DITTO F1	DeepMatcher F1	Magellan F1
Structured	iTunes-Amazon	97.0	88.5 +8.5	91.2 +5.8
	DBLP-ACM	99.0	98.4 +0.6	98.4 +0.6
	DBLP-Scholar	95.6	92.3 +3.3	94.7 +0.9
	Walmart-Amazon	86.8	66.9 +19.9	71.9 +14.9
Textual	Abt-Buy	89.3	62.8 +26.5	43.6 +45.7
	Amazon-Google	75.6	69.3 +6.3	49.1 +26.5
	WDC Computer - Large	91.7	89.5 +3.2	64.5 +27.2
	WDC Computer - Small	80.8	70.5 +10.3	57.6 +23.2

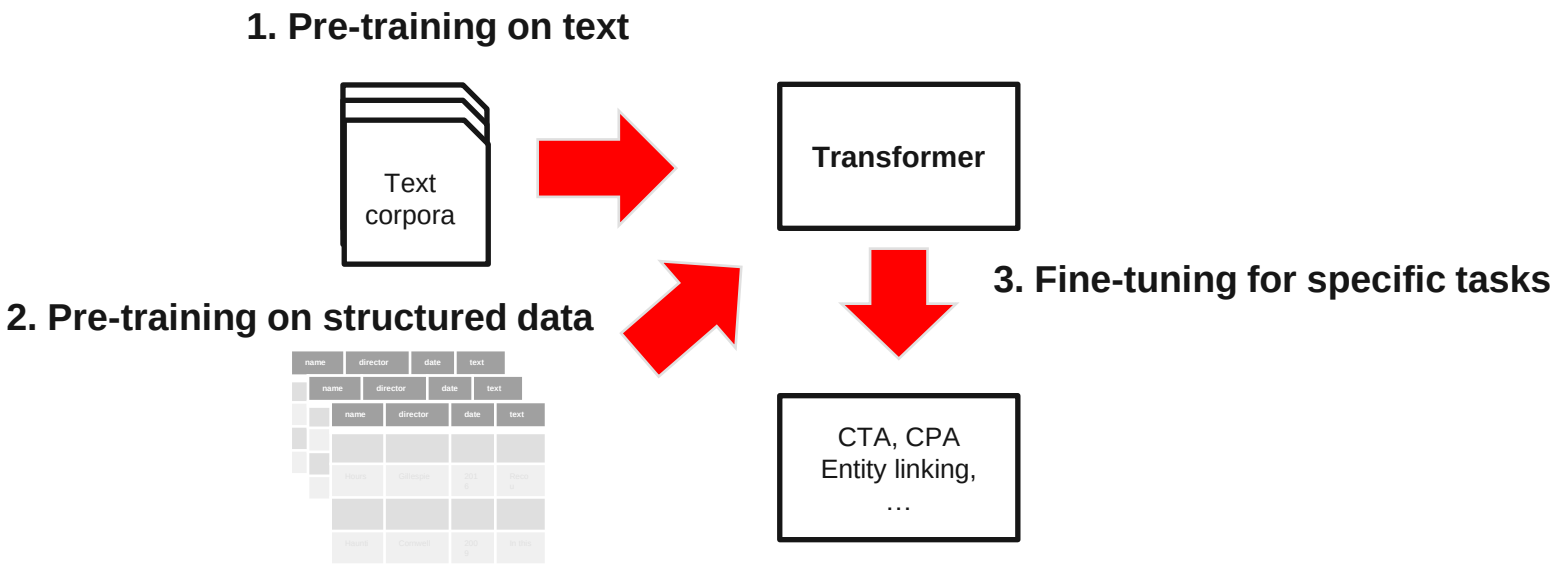
- constant improvement for structured data
- large performance gain for textual data

Potential Reasons for the Performance Gain

- Serialization allows to pay attention to all attributes
  - no strict separation between attributes
- WordPiece tokenizer breaks unknown terms into pieces
  - no problems with out of vocabulary terms
- Transfer learning from pre-training texts
  - different surface forms are already close in embedding space
- Contextualization of the embeddings
  - potentially more suited for capturing differing semantics

Transformers for Table Annotation

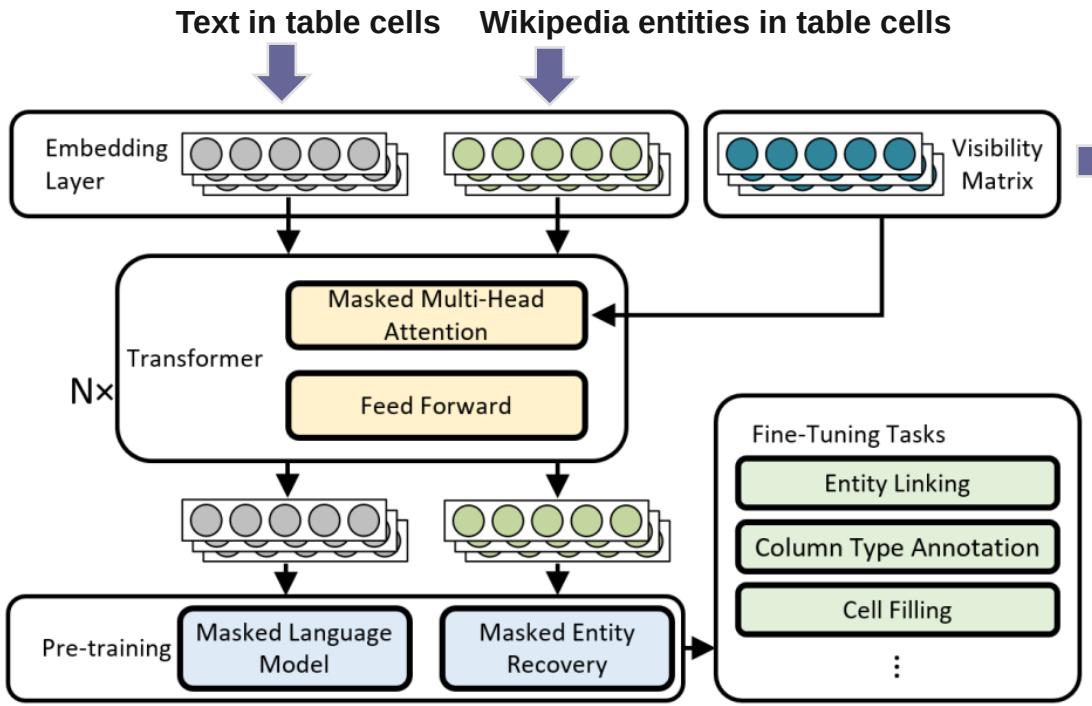
- **TURL** (http://www.vldb.org/pvldb/vol14/p307-deng.pdf)
  - uses TinyBERT as base architecture, pre-trained on text corpora
  - Additional pre-training on ~600k Wikipedia tables using **Masked Language Modeling** and a **Masked Entity Recovery** objective
  - Fine-tuned and evaluated on entity linking (CEA), CTA, CPA, row population and cell filling



Deng, et al.: **TURL: Table Understanding through Representation Learning**. PVLDB 2020.

Transformers for Table Annotation

– **Visibility Matrix** = attention mask during self-attention calculation
  - Determines which other tokens from the other table cells are visible for each token



(a) Visibility for *caption tokens*

name	director	date	description

(b) Visibility for header tokens

name	director	date	description
	Michael Tully		
	Craig Gillespie		
	James Strong		
	Peter Cornwell		

(c) Visibility for cell entities

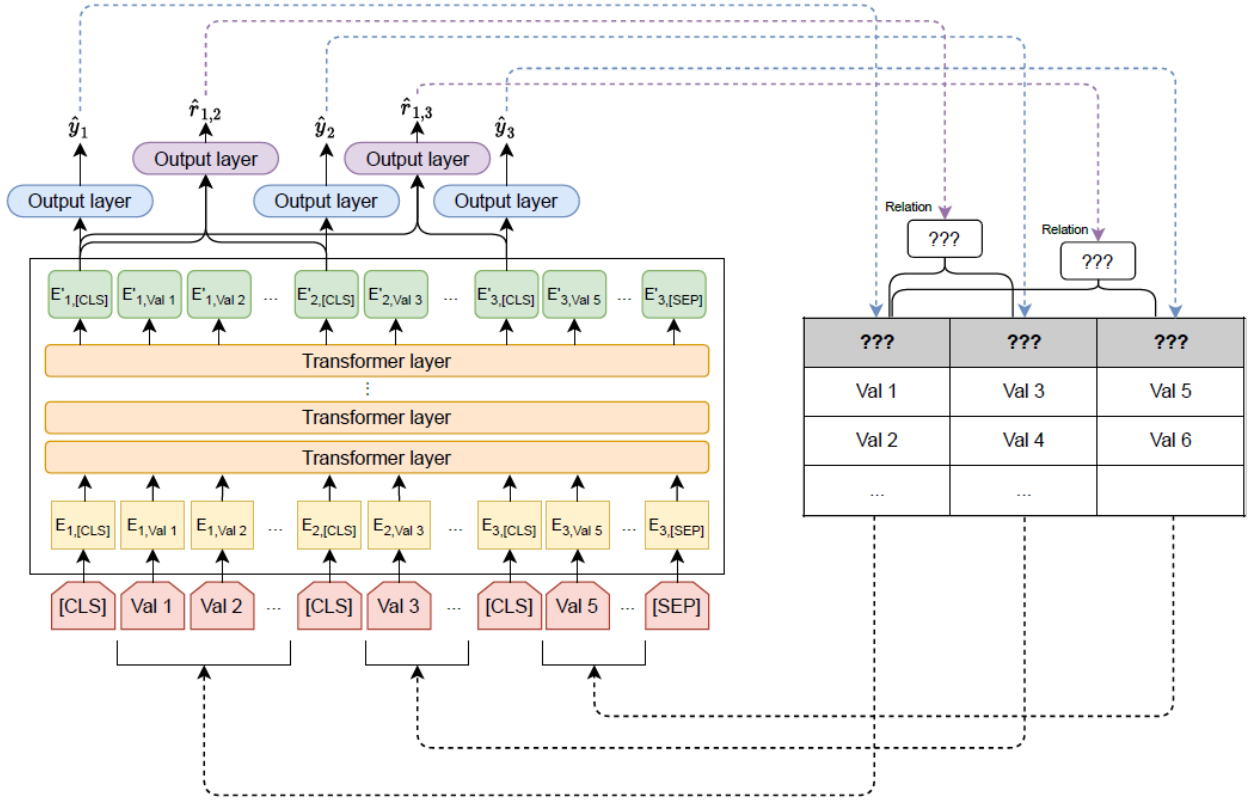
	director		
	Michael Tully		
	Craig Gillespie		
United	James Strong	2011	A devastating …
	Peter Cornwell		

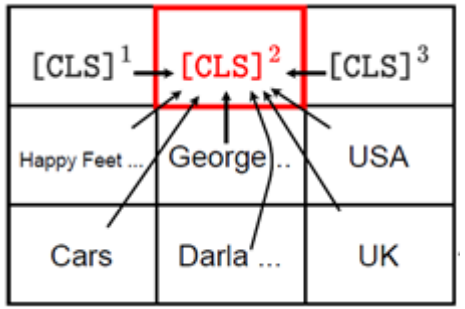
Deng, et al.: **TURL: Table Understanding through Representation Learning**. PVLDB 2020.

Transformers for Table Annotation

– **DoDuo** (https://arxiv.org/abs/2104.01785)

- fine-tunes BERT (no further pre-training) for CTA and CPA and uses multi-task learning
- evaluation of single-column and multi-column models





Suhara, et al.: **Annotating Columns with Pre-trained Language Models**. SIGMOD 2022.

Evaluation Results of Table Annotation Systems on WikiTables

- Column Type Annotation (~255 types)

Method	F1	P	R
TURL (TinyBERT)	88.86	90.54	87.23
DoDuo (BERT)	92.45	92.45	92.21

- Column Property Annotation (~121 relations)

Method	F1	P	R
TURL (TinyBERT)	90.94	91.18	90.69
DoDuo (BERT)	91.72	91.97	91.47

Suhara, et al.: Annotating columns with pre-trained language models. SIGMOD, 2022.
Deng, et al.: TURL: table understanding through representation learning. PVLDB 2020.

Further Table Embedding Approaches

- **TABBIE** (https://aclanthology.org/2021.naacl-main.270.pdf)
  - general pre-training on 1.8M Wikipedia Tables and 24.8M WebTables
  - using a cell corruption objective: Replace some cell contents with frequency-based cell sampling across all tables and then try to predict if a cell was changed or not
  - evaluated on Column Type Annotation, Row Population, Column Population
  - code available
- **TUTA** (https://dl.acm.org/doi/abs/10.1145/3447548.3467434)
  - pre-trained on relational and spreadsheet tables from Wikipedia and the WDC WebTable Corpus
  - using a Masked Language Model, multi-choice Cloze at the cell level and context retrieval at table level
  - evaluated on Cell Type and Table Type Classification
  - code not available

More at: Pujara, et al**.: From Tables to Knowledge: Recent Advances in Table Understanding**. Tutorial at KDD2021.

https://usc-isi-i2.github.io/KDD21Tutorial/

Evaluation Campain for Table to Knowledge Graph Matching

- Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)
- Yearly challenge where table annotation methods are compared on their performance regarding the **CTA**, **CPA** and **CEA** tasks on the same benchmark datasets
- Table columns and cells are linked to a Knowledge Graph (such as DBpedia or WikiData) concept

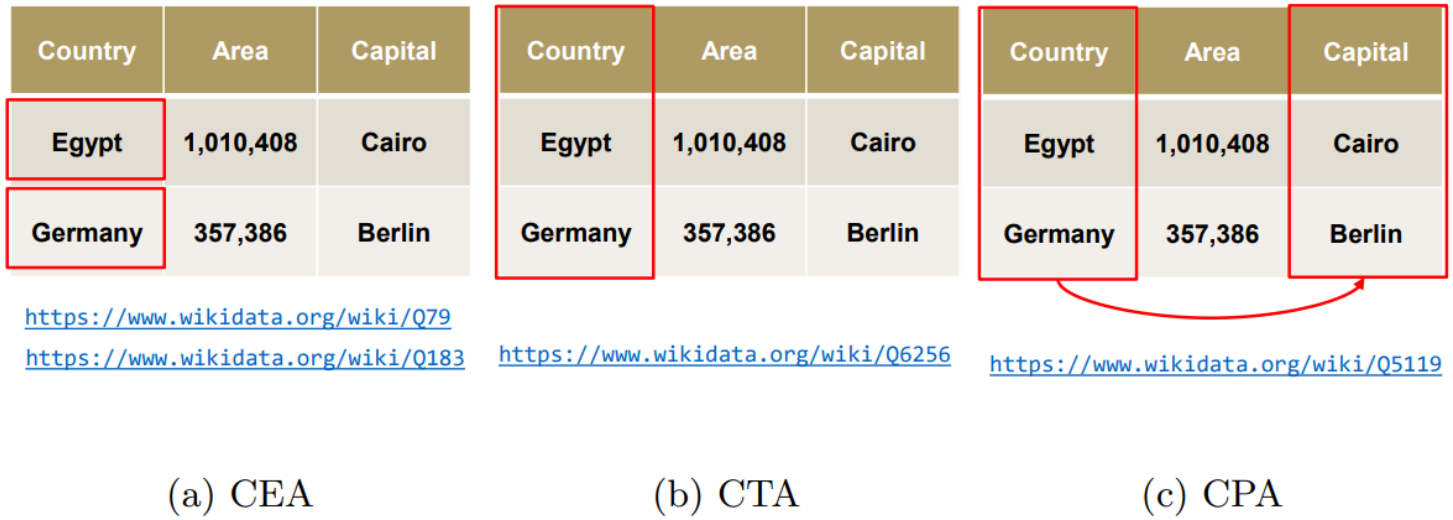


Image Source: N. Abdelmageed, et al.: BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data

E. Jiménez-Ruiz et al.: **SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems.** ISWC 2020.
http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

Benchmarks using Annotated Tables as Training Data

- **WikiTables**
  - 580,171 tables extracted from Wikipedia
  - 406,706 of these tables are labled for the Column Type Annotation (CTA) task (255 labels)
  - 55,970 tables labeled for the Columns Property Annotation (CPA) task (121 labels)
  - labels are taken from Freebase knowledge graph
  - Task: Predict the CTA and CPA labels for tables where the true labels are hidden
- **Schema.org Table Annotation Benchmark (SOTAB)**
  - tables containing schema.org data originating from 74,215 different websites
  - 59,548 tables annotated for CTA (91 labels)
  - 48,379 tables annotated for CPA (176 labels)
  - labels are mostly taken from the schema.org vocabulary
  - Task: Predict the CTA and CPA labels for tables where the true labels are hidden

Project Goals

1. Try to improve the state-of-the-art performance for the CTA and CPA tasks on the SOTAB benchmark, by experimenting with
   1. different table serialization techniques
   2. table augmentation techniques
   3. different methods for embedding table columns plus context
2. Compare performance of your methods to
   - existing table annotation methods (e.g DoDuo, TURL, …) using different evaluation datasets
3. Try to explain why models perform better than others
   - evaluate performance on different challenge columns
   - conduct an error analysis

Learning Targets

Improve your technical skills

- Work as a **Data Scientist**: clean, profile, classify data
- Understand the nature of **Web Data**
- Improve your technical expertise concerning **Deep Learning**
- Improve your programming skills

Improve your soft skills

- Work as part of a bigger team on a more complex project
- Organize yourself and assign tasks based on your skills
- Communicate and coordinate your work

The WDC Schema.Org Table Annotation Benchmark

1. Semantic Annotations in HTML Pages
2. Web Data Commons – Schema.org Table Corpus
3. Schema.org Table Annotation Benchmark (SOTAB)

Semantic Annotation of HTML Pages: Schema.org

- 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

Example: Microdata Annotations in HTML


```
<div  itemtype="http://schema.org/Product">
  <span itemprop="name">Sony GTK-XB5L Audiosystem</span>
  <span itemprop=“gtin13“>04048945021687</span>
  <span itemprop=“description">high-power home audio system with Bluetooth technology</span>
</div>
 <div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
    <span itemprop="ratingValue"> 4 </span> stars-based on
    <span itemprop="reviewCount"> 250 </span> reviews.
 </div>
```

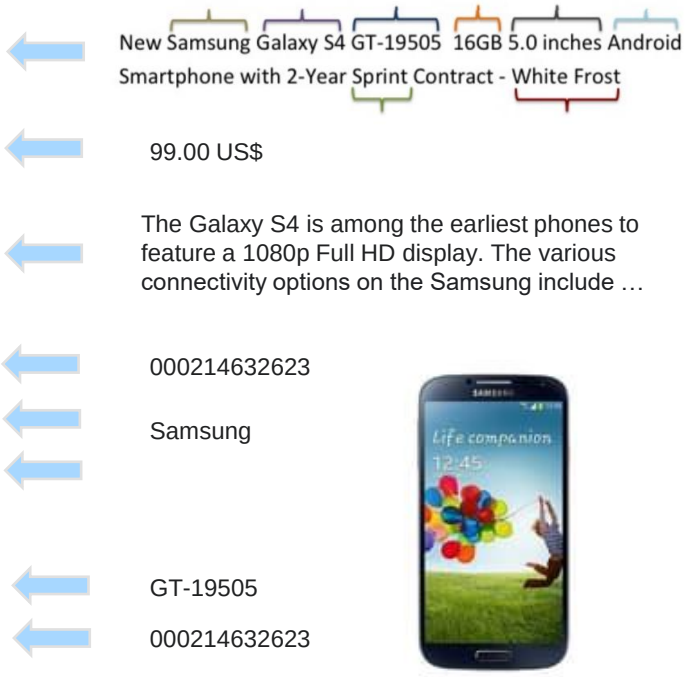
Frequently used Schema.org Classes (2020)

Class	# Websites (PLDs)	
	JSON-LD	Microdata
schema:WebPage	4,484,026	1,339,999
schema:Person	3,151,809	514,990
schema:BreadcrumbList	1,688,820	924,991
schema:Article	1,327,578	627,303
schema:Product	1,234,972	1,059,149
schema:Offer	1,182,855	946,725
schema:PostalAddress	863,243	585,417
schema:BlogPosting	529,020	552,338
schema:LocalBusiness	363,843	280,338
schema:AggregateRating	432,014	315,253
schema:Place	255,139	93,124
schema:Event	194,115	77,722
schema:Review	181,097	158,333
schema:JobPosting	28,759	8,520

http://webdatacommons.org/structureddata/2020-12/stats/schema_org_subsets.html

Properties used to Describe Products (2020)

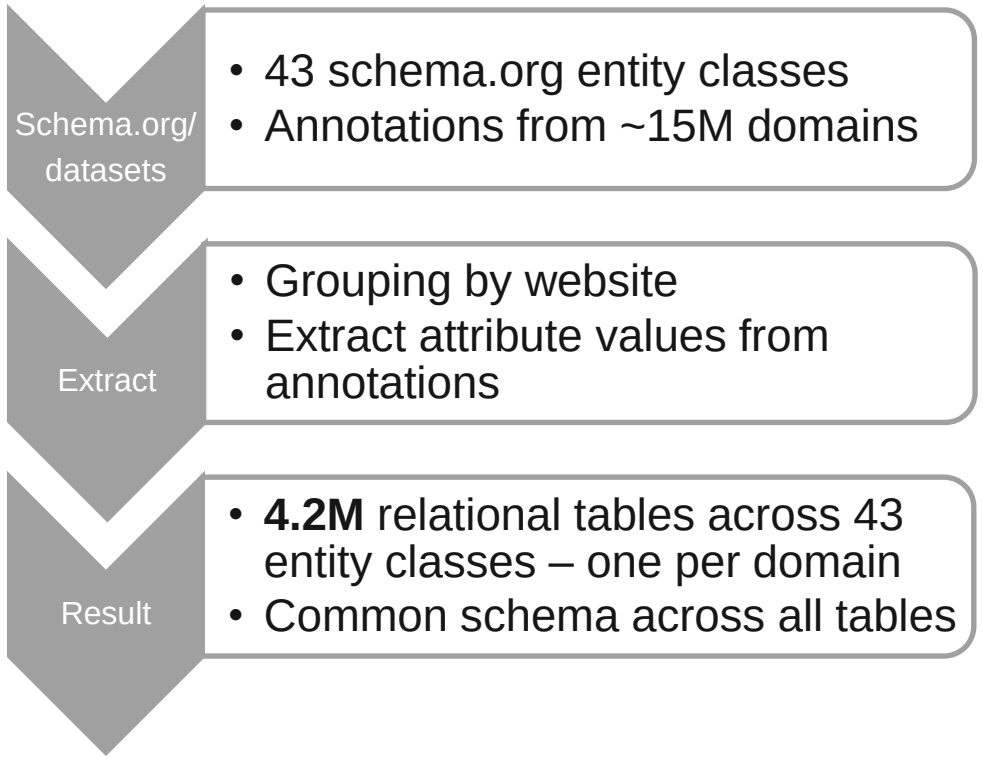
Attribute	% of PLDs
schema:Product/name	99 %
schema:Product/offers	94 %
schema:Offer/price	95 %
schema:Offer/priceCurrency	95 %
schema:Product/description	84 %
schema:Offer/availability	72 %
schema:Product/sku	56 %
schema:Product/brand	30 %
schema:Product/image	26 %
schema:Product/aggregateRating	17 %
schema:Product/mpn	6.3 %
schema:Product/productID	4.7 %
…	…



http://webdatacommons.org/structureddata/schemaorgtables/

The WDC Schema.Org Table Corpus

- Extract annotations for 43 schema.org entity classes
- Extract attribute values and group entities by website
- Initial cleaning steps are performed
- **Final result**: 4.2M tables - one table per domain, containing all annotated entities after cleaning.
- All tables share the same schema



Details and Download:

- **http://webdatacommons.org/structureddata/schemaorgtables/**

The WDC Schema.org Table Annotation Benchmark

- covers the Column Type Annotation and Column Property Annotation tasks
- uses tables from the in Schema.org Table Corpus
  - tables with schema.org labels are given for training
  - tables with masked-out labels are used for testing
- Table columns selected based on three challenges:
  - **Missing values:** Columns where missing cells are present, between 10-70 % density
  - **Value Format Heterogeneity:** Columns that contain values with different (measurement) formats
  - **Corner Cases:** Columns that are hard to annotate, similar columns that have different labels, dissimilar columns that have the same label
  - **Random columns:** Randomly chosen columns for all chosen labels

→ **Result:** ~135,000 annotated columns and column relations from ~50,000 tables for Column Type Annotation and Column Property Annotation tasks separately

Benchmark Statistics

– **Label Space:** 91 semantic types for CTA and 176 relation labels for CPA.

  • Corresponding to 17 domains: Product, JobPosting, Hotel, Restaurant, Event, MusicAlbum, Person ...

CPA Overall Top 10 Labels/Annotated Columns

description	7824
datePublished	4963
telephone	4731
author	4151
price	4132
startDate	3918
priceCurrency	3597
ratingValue	3546
addressCountry	3373

CTA Overall Top 10 Labels/Annotated Columns

Text	7771
DateTime	6428
Duration	5997
Mass	5649
Date	5508
currency	5383
Number	5160
telephone	4647
price	4475

Benchmark Statistics

- Tables have a minimum of 10 rows and a minimum of 3 columns.
- Includes textual, numerical and datetime column values
- Fixed training, validation and testing splits
  - Training set is provided in two sizes: **Large** training set and **Small** training set

Overall Top 5 Table Types in SOTAB

	Column Type Annotation											Columns Property Annotation										
	Overall			Training Set		Small Training Set		Validation Set		Testing Set		Overall			Training Set		Small Training Set		Validation Set		Testing Set	
Schema.org Type	# tables	# columns	median rows/cols	# tables	# columns	# tables	# columns	# tables	# columns	# tables	# columns	# tables	# columns	median rows/cols	# tables	# columns	# tables	# columns	# tables	# columns	# tables	# columns
Overall	59,548	162,351	33/7	46,790	130,471	11,517	33,004	5,732	16,840	7,026	15,040	48,379	174,998	42/8	37,128	134,425	9,435	33,643	4,771	17,417	6,480	23,156
Product	12,534	30,856	40/8	10,403	25,353	2,037	5,211	857	2,507	1,274	2,996	16,032	36,464	65/10	12,668	29,149	3,075	6,378	1,476	2,726	1,888	4,589
Event	6,206	17,972	23/7	4,721	14,150	1,386	4,631	593	1,869	892	1,953	6,713	21,298	25/8	5,220	15,949	1,417	4,868	722	2,573	771	2,776
LocalBusiness	5,292	14,941	37/9	3,744	10,657	1,226	3,932	602	2,161	946	2,123	5,274	15,987	39/9	4,076	11,373	1,108	3,615	503	1,991	695	2,623
Recipe	5,080	24,590	42/13	4,083	20,809	756	4,459	412	2,199	585	1,582	4,535	27,070	52/15	3,547	20,279	967	4,687	426	2,426	562	4,365
CreativeWork	4,674	9,240	26/4	3,393	6,921	1,142	2,299	686	1,355	595	964	2,997	12,468	28/6	2,323	9,866	584	2,681	316	1,360	358	1,242
Person	4,623	10,466	27/5	3,818	8,680	851	1,876	374	867	431	919	3,186	11,535	32/5	2,159	8,542	476	1,936	309	1,033	718	1,960

Ground Truth Files

- Tables and ground truth files provided
  - **Tables** provided without column names

	0	1	2	3	4	5	6
0	Designer Inspired Chain Necklace Glass Dome Pe...	40\" Chain Classic designer inspired famous fl...	USD	69	https://schema.org/InStock	Beauty In Stone Jewelry	2020-10-07
1	Genuine Pearl With Modern Cube on Chain Necklace	Bridal Pearl Necklace Genuine freshwater pearl...	USD	49	https://schema.org/InStock	Beauty In Stone Jewelry	2020-10-07
2	Beach Glass Necklace With Crystal Pendant	Dark gray beach glass, metal patina beads on n...	USD	99	https://schema.org/InStock	Beauty In Stone Jewelry	2020-10-09
3	Brushed Gold Bracelet	22k Gold Electroplate over brass beads. Stretc...	USD	65	https://schema.org/InStock	Beauty In Stone Jewelry	2020-10-03

  - **Ground Truth files** for training, validation and 5 test sets: full test set, missing values test set, random test set, format heterogeneity test set and corner cases test set.

	table_name	column_index	label
0	Product_beautyinstonejewelry.com_September2020...	3	price
1	Product_beautyinstonejewelry.com_September2020...	2	currency
2	Product_beautyinstonejewelry.com_September2020...	5	Brand

CTA Ground Truth Example

	table_name	main_column_index	column_index	label
Product_corememoriesco.com_September2020_CPA.j...	0	1	url
Product_corememoriesco.com_September2020_CPA.j...	0	2	priceCurrency
Product_corememoriesco.com_September2020_CPA.j...	0	5	image

CPA Ground Truth Example

Team Project Organization

Duration: 6 months (03.10.2022 – 03.04.2023)

Participants: 5 people

Type of work: Team and subgroup based

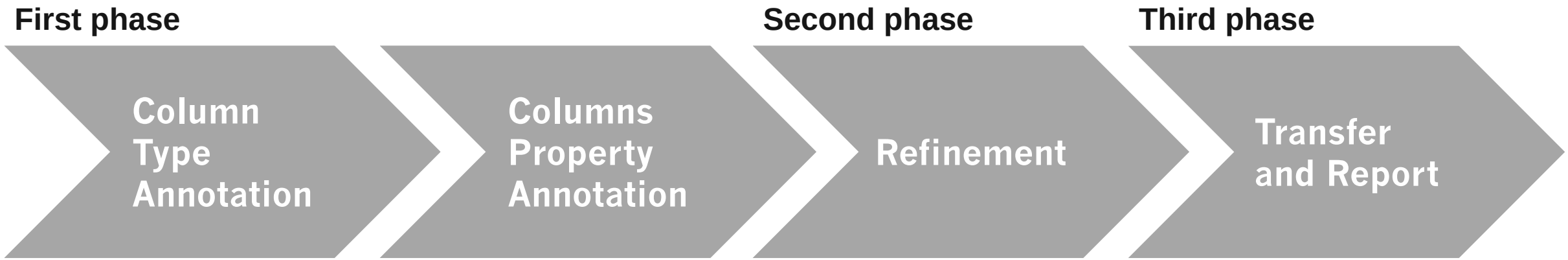
Milestones: 3 project phases

ECTS Points: 12

Evaluation

- Intermediate presentations
- Final report as HTML page
- Individual contribution to the deliverables is graded

Team Project Organization: Project Phases



Phase 1a: Column Type Annotation

Participants: all team members

Duration: 3.10 – 14.11

Input: SOTAB CTA dataset (*provided here*)

Sub-phases:

1. Topic understanding and ideas collection
    - Read papers, get ideas about **Table Augmentation** and **Table Serialization** and decide which methods to implement and try out
    - **Sub-phase duration:** 3.10 – 10.10
    - **Deliverable:** List of methods and description of each method

Table Serialization: Ideas where to start

- What is table serialization?
  - Language models take **token sequences** as input
  - **Table Tasks:** columns are converted into token sequences
- **Basic Serialization:** Serialize one column
  - Concatenate all column values
  - Add special [CLS] token at the beginning of sequence
  - Add special [SEP] token at the end of the sequence

[CLS] Alofi Nice Dublin [SEP]

Lau's Gateway	209 Main Street	Alofi	NU	NZD
Radisson Blu Hotel, Nice	223 Promenade Des Anglais	Nice	FR	EUR
Phoenix Park Hotel	38-39 Parkgate Street Dublin 8	Dublin	IE	EUR

- **Table Serialization:** Serialize all table columns
  - Concatenate all column values and separate them with [CLS] tokens, add a [SEP] token at the end: [CLS] Lau‘s Gateway Radisson ... [CLS] 209 Main Street ... [CLS] Alofi Nice .... [CLS] NZD ... [SEP]
- Think of **other serialization possibilities** that could be used to improve models:
  - Concatenate only neighboring columns or break down tables into smaller tables
  - Summarize long column values using ex. TF-IDF, if table is long sampling from values

Table Augmentation: Ideas where to start

- Try data augmentation techniques in the context of tables
- At different levels:
  - **Cell level**
  - **Row/column level**
  - **Table level**
- Examples:
  - **Cell level:** Delete cell values, Replace tokens ...
  - **Row/column level:** Shuffle column value or row value order
  - **Table level:** Mix rows and columns from different tables
- Can table augmentation improve performance results, especially using the small training set where less examples are available?
- You can get further ideas in this tutorial for data augmentation in matching tasks: https://vldb.org/2021/files/slides/tutorial/tutorial7.pdf

Phase 1a: Column Type Annotation

Participants: all team members

Duration: 3.10 – 14.11

Input: SOTAB CTA dataset (*provided here*)

Sub-phases:

2. Implement methods and run experiments
   - Baseline methods (Random Forest, SVM), BERT, RoBERTa, Doduo, Contrastive Learning
   - **Sub-phase duration:** 10.10 – 07.11
   - **Deliverables:** Code and Evaluation results
3. Analyze results and **develop ideas for** improvement for **second iteration**
   - Error analysis: Look at correctly/incorrectly classified examples for select models and calculate statistics for error classes
   - **Sub-phase duration:** 07.11 – 14.11
   - **Deliverables:** Error analysis report and plans for the next iteration

Phase 1a: Column Type Annotation - How to get started?

- Get the Tables (here)
- Get acquainted with data, ground truth files, etc.
- Experiment with table augmentation and serialization using baseline methods such as BERT, or more advanced methods like DODUO (https://github.com/megagonlabs/doduo)
- Code to get started will be provided in Github in a shared repository
- We will provide access to GPU server in the DWS student server
- You can also use the BW Uni Cluster
  - https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0

Phase 1b: Columns Property Annotation

Participants: all team members

Duration: 14.11 – 19.12

Input: SOTAB CPA dataset (*provided here*)

Sub-phases:

1. Topic understanding and ideas collection
    - Read papers, get ideas about **Table Augmentation** and **Table Serialization** and decide what methods to use in the evaluation phase
    - **Sub-phase duration:** 14.11 – 21.11
    - **Deliverables:** List of ideas to try and description

Table Augmentation and Serialization: Ideas for CPA

- **Table Augmentation for CPA:** focus on augmention of pairs of columns within the same table
- Some similar examples as in the CTA phase:
  - **Cell level:** Delete cell values, Replace tokens ...
  - **Row/column level:** Shuffle column value or row value order
- **Table Serialization for CPA:** Normally treats a column pair as a sequence
  - Concatenate all column values of each column
  - Add [CLS] token at the beginning and [SEP] token to separate the two columns and at the end of the sequence
  - Further idea: Add other columns as context
  - Summarize long textual columns (TF-IDF method)

Phase 1b: Columns Property Annotation

Participants: all team members

Duration: 14.11 – 19.12

Input: SOTAB CPA dataset (*provided here*)

Sub-phases:

2. Implement methods and run experiments
   - Baseline methods (Random Forest, SVM), BERT, RoBERTa, Doduo, Contrastive Learning
   - **Sub-phase duration:** 21.11 – 12.12
   - **Deliverables:** Code and Evaluation results
3. Analyze results and **develop ideas for** improvement for **second iteration**
   - Error analysis: Look at correctly/incorrectly classified examples for select models and calculate statistics for error classes
   - **Sub-phase duration:** 12.12 – 19.12
   - **Deliverables:** Error analysis report and plans for the next iteration

Phase 2: Refinement (Second Iteration)

Participants: two subgroups (one subgroup per CTA/CPA task)

Duration: 19.12 – 06.02

Sub-phases:

1. Improve code for CTA and CPA task and rerun experiments
   - Based on the plan for refinement made on the previous phase
   - **Sub-phase duration:** 19.12 – 16.01
   - **Deliverables:** Code and Evaluation Results
2. Error analysis
   - Look at correctly/incorrectly classified examples for select models
   - How did the models perform in challenge columns (Missing values...) and in random columns?
   - How did the models perform on different types of data (numerical, datetime, text)?
   - **Sub-phase duration:** 16.01 – 06.02
   - **Deliverables:** Error Analysis Report

Phase 3: Transfer and Report

Participants: all team members or two subgroups per task (CPA/CTA)

Duration: 06.02 – 03.04

Input: Results from Phase 2

Sub-phases:

1. Test good working approaches on other datasets
   - such as **Wikitables** by TURL and **GitTables** benchmark
   - include results from error analysis into SOTAB test and training
   - **sub-phase duration:** 06.02 – 06.03
2. Write report
   - as an HTML page
   - sample HTML code will be provided
   - **sub-phase duration:** 06.03 – 03.04

Schedule

Date	Session
Wednesday, 28.09.2022	Kickoff meeting (today)
	Phase 1a: CTA Topic Understanding and ideas collection
Monday, 10.10.2022	1st Deliverable: List of methods and description of each method
	Phase 1a: CTA Experiments
Monday, 21.10.2022	Meet Keti and report current plan and results
Monday, 07.11.2022	2nd Deliverable: Code and evaluation results
	Phase 1a: CTA improvements plan
Monday, 14.11.2022	3rd Deliverable: Error analysis report and plans for the next iteration
	Phase 1b: CPA Topic Understanding and ideas collection
Monday, 21.11.2022	4th Deliverable: List of methods and description of each method
	Phase 1b: CPA Experiments
Monday 28.11.2022	Meet Keti and report current plan and results
Monday, 12.12.2022	5th Deliverable: Code and evaluation results
	Phase 1b: CPA improvements plan
Monday, 19.12.2022	**6th Deliverable: Intermediate presentation to Professor Bizer, code & data**
	Phase 2 (in 2 subgroups): Refinement
Monday, 16.01.2023	7th Deliverable: Evaluation Results
	Phase 2 (in 2 subgroups): Error Analysis

Schedule

Date	Session
Monday, 16.01.2023	Meet Keti and report current plan and results
Monday, 06.02.2023	8th Deliverable: Error Analysis Report
	Phase 3: Transfer
Monday, 20.02.2023	Meet with Keti and discuss plans and results
Monday, 06.03.2023	**Final Presentation to Professor Bizer**
	Phase 3: Write Report as HTML page
Monday, 03.04.2023	**Submission deadline**

Formal Requirements & Consultation

Deliverables

1. On the deliverable dates provide us via e-mail with:
   - **Presentation slides**
   - **Task to member report**: excel sheet stating which team member conducted which subtask
   - **Code/Data**: link or zipped folder with your code and data

2. **Final Report as HTML-page**
   - **Giving overview of project results.**
   - **Providing artefacts/code, see team project**

All deliverables should be sent to Keti & Chris!

Formal Requirements & Consultation

Final grade

- 20% for each phase (CTA, CPA, Refinement, and Transfer)
- 20% for final report
- Late submission: -0.3 per day

Consultation

- Send one e-mail per team or subgroup stating your questions to Keti

Useful Software

- Transformers code
  - Code to start in shared repository
  - HuggingFace Transformers: https://huggingface.co/transformers/
  - TURL: https://github.com/sunlab-osu/TURL
  - DODUO: https://github.com/megagonlabs/doduo
- Processing and GPUs
  - Teaching GPU-Server
  - Google Colab: https://colab.research.google.com/
  - BwUniCluster2.0: https://wiki.bwhpc.de/e/Category:BwUniCluster_2.0
- Team Cooperation
  - GitHub/Lab for the code base
  - Project Management Tool of your choice

Related Work: (Tabular) Transformers (1/2)

- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, "**TURL: table understanding through representation learning**," Proc. VLDB Endow., vol. 14, no. 3, pp. 307–319, Nov. 2020.
- Y. Suhara et al., "**Annotating Columns with Pre-trained Language Models**," arXiv:2104.01785 [cs], Apr. 2021
- H. Iida, D. Thai, V. Manjunatha, and M. Iyyer, "**TABBIE: Pretrained Representations of Tabular Data**," arXiv:2105.02584 [cs], May 2021
- N. Tang et al., "**RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation**," arXiv:2012.02469 [cs]
- P. Yin, G. Neubig, W. Yih, and S. Riedel, "**TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data**," arXiv:2005.08314 [cs], May 2020
- J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. Eisenschlos, "**TaPas: Weakly Supervised Table Parsing via Pre-training**," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, Jul. 2020, pp. 4320–4333.
- D. Wang, P. Shiralkar, C. Lockard, B. Huang, X. L. Dong, and M. Jiang, "**TCN: Table Convolutional Network for Web Table Interpretation**," arXiv:2102.09460 [cs], Feb. 2021

Related Work: (Tabular) Transformers (2/2)

- Z. Wang et al., "**TUTA: Tree-based Transformers for Generally Structured Table Pre-training**," in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, New York, NY, USA, Aug. 2021, pp. 1780–1790.
- Badaro, G., Saeed, M., & Papotti, P. (2021). **Transformers for Tabular Data Representation: A Survey of Models and Applications.** EURECOM, 2021, technical Report.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. **Attention Is All You Need**. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 6000–6010.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.** In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 4171–4186.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, et al. 2019**. RoBERTa: A Robustly Optimized BERT Pretraining Approach**. arXiv:1907.11692 (2019).

Related Work: General (Deep) Schema Matching

- M. Hulsebos et al., “**Sherlock: A Deep Learning Approach to Semantic Data Type Detection**,” arXiv:1905.10688 [cs, stat], May 2019
- Zhang, D., Suhara, Y., Li, J., Hulsebos, M., Demiralp, Ç. and Tan, W.C. **Sato: Contextual semantic type detection in tables**. arXiv preprint arXiv:1911.06311. 2019
- J. Chen, E. Jimenez-Ruiz, I. Horrocks, and C. Sutton, “**ColNet: Embedding the Semantics of Web Tables for Column Type Prediction**,” arXiv:1811.01304 [cs]
- J. Chen, E. Jimenez-Ruiz, I. Horrocks, and C. Sutton, “**Learning Semantic Annotations for Tabular Data**”, in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, Aug. 2019, pp. 2088–2094
- J. Zhang, B. Shin, J. D. Choi, and J. C. Ho, “**SMAT: An Attention-Based Deep Learning Solution to the Automation of Schema Matching**,” in Advances in Databases and Information Systems, Cham, 2021, pp. 260–274

Related Work: Table Annotation Benchmarks

- M. Hulsebos, Ç. Demiralp, P. Groth, **GitTables: A Large-Scale Corpus of Relational Tables**, arXiv:2106.07258 (2022).
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, "**TURL: table understanding through representation learning**," Proc. VLDB Endow., vol. 14, no. 3, pp. 307–319, Nov. 2020.
- N. Abdelmageed, S. Schindler, B. König-Ries, **BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data**, in: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, volume 3103, pp. 13–18, 2021
- Cutrona, V., Bianchi, F., Jiménez-Ruiz, E., & Palmonari, M. **Tough tables: Carefully evaluating entity linking for tabular data**. In International Semantic Web Conference, pp. 328-343, Nov. 2020
- E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, **SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems**, in: The Semantic Web, Springer International Publishing, pp. 514–530, 2020
- D. Ritze, C. Bizer, **Matching Web Tables To DBpedia - A Feature Utility Study**, in: Proceedings of the 20th International Conference on Extending Database Technology, pp. 210–221, 201
- More related work at PapersWithCode: https://paperswithcode.com/task/column-type-annotation, https://paperswithcode.com/task/columns-property-annotation

Questions?