JenTab: A Toolkit for Semantic Table Annotations

Nora Abdelmageed, Sirko Schindler

Friedrich Schiller University Jena, Germany

ESWC 2021

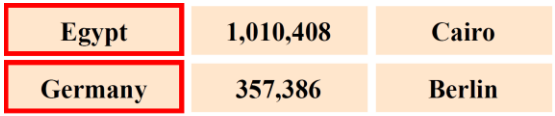
FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA Fakultät für Mathematik
und Informatik

FUSION

MichaelStifelCenterJena
for Data-Driven and Simulation Science

ESWC21

Semantic Table Annotation Tasks

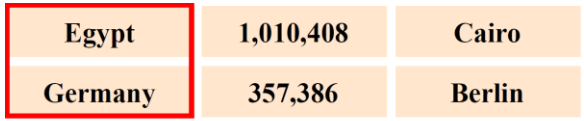


Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wd:Q79 ("Egypt")
wd:Q183 ("Germany")

CEA

Cell **E**ntity **A**nnotation

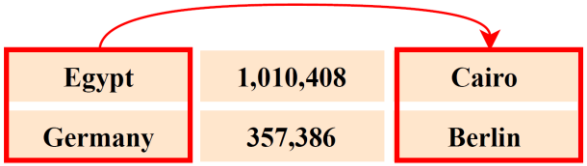


Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wd:Q6256 ("country")

CTA

Column **T**ype **A**nnotation



Egypt	1,010,408	Cairo
Germany	357,386	Berlin

wdt:P36 ("capital")

CPA

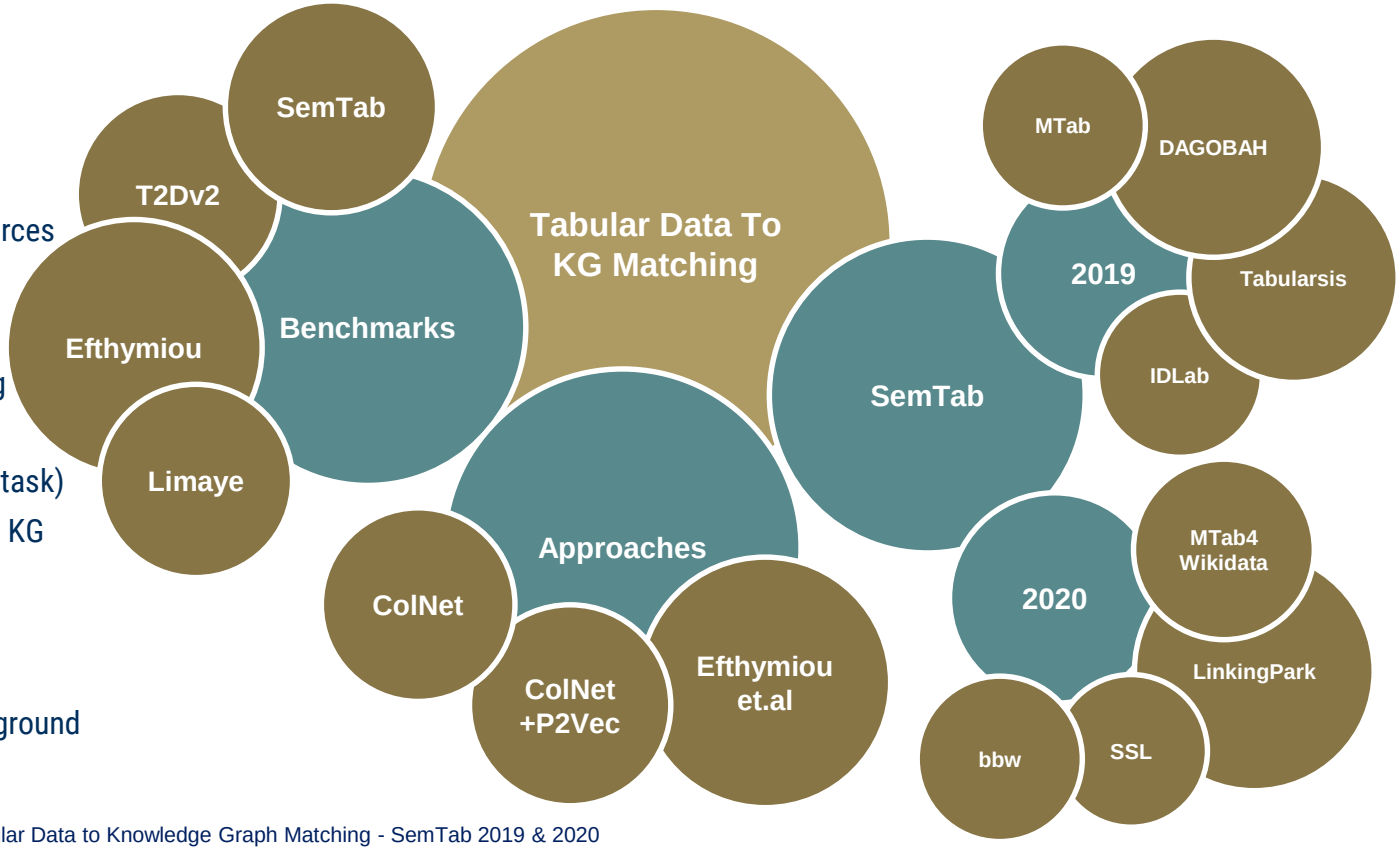
Column **P**roperty **A**nnotation

Outlook

1. Semantic Table Annotations
2. Related Work
3. Proposed Technique
4. Datasets
5. Evaluation
6. Results
7. Conclusions & Future work

Related Work

- **SemTab***
  - Multiple data sources
  - Full cell query
- **Approaches**
  - Machine Learning
  - Semantics
  - Limited scope (1 task)
  - Cannot cope with KG changes
- **Benchmarks**
  - Small
  - Some have poor ground truth (Limaye)



* Knowledge Semantic Web Challenge on Tabular Data to Knowledge Graph Matching - SemTab 2019 & 2020

Proposed Technique

- Preprocessing
- CFS Pattern
- Contexts and annotation modules
- Architecture

Preprocessing

1. **Generic fix**
   - Encoding fixes (ftfy)
   - Special character removal
   - Restore missing spaces (parse errors)
2. **Datatype predication**
   - Types with equivalents in KG
   - `OBJECT, QUANTITY, DATE, STRING`
3. **Type-based cleaning**
   - Extract the relevant part from (`QUANTITY, DATE`)
   - 10/12/2020 (10 Dec 2020) → 2020-12-10
   - 1,199 km (745 mi) → 1199

CFS Pattern

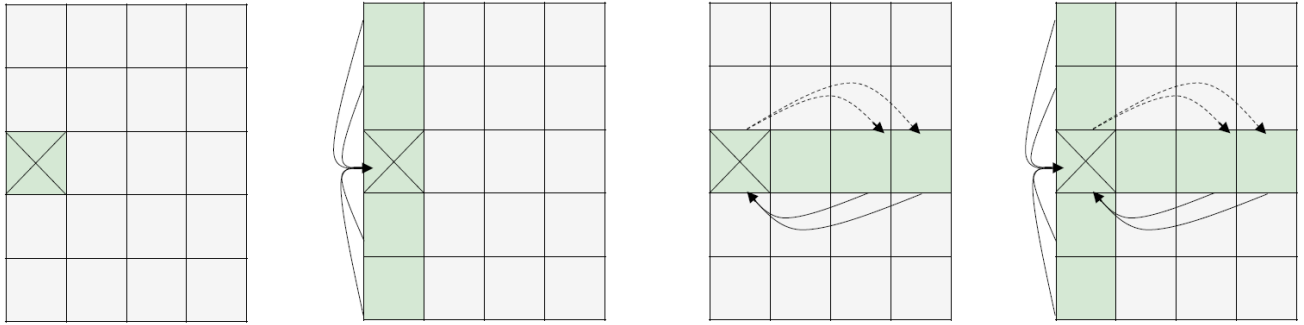
- **Different KG-lookups using different constraints, assemble all the retrieved information and find a proper solution.**

- **C**reate, **F**ilter and **S**elect pattern
  1. **Create** all possible candidates for the 3 tasks
  2. **Filter** the initial candidates set by using feedback from other tasks
  3. **Select** the most suitable candidate from the remaining ones

- Data sources
  - Lookup services (elastic search tools over KG)
  - SPARQL endpoint (structured queries)

Contexts and Annotation Modules

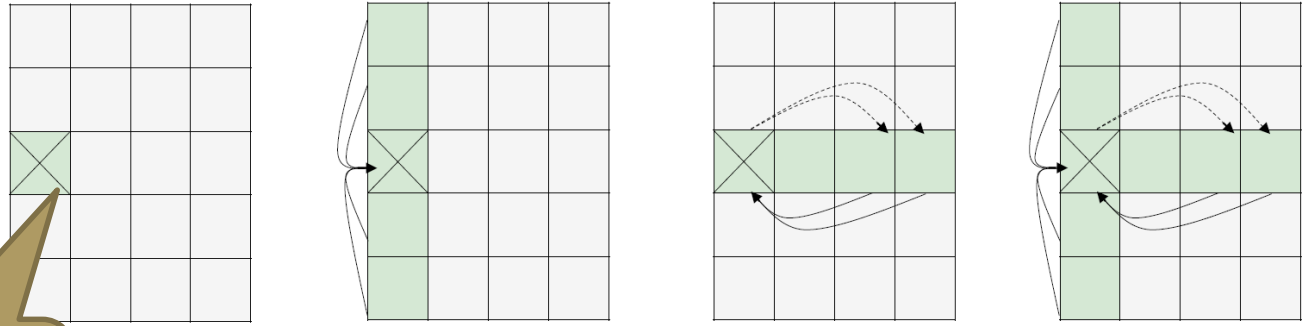
- 4 contexts used to create/filter annotations



(a) Cell (b) Column (c) Row (d) Row-Column

Contexts and Annotation Modules

- 4 contexts used to create/filter annotations

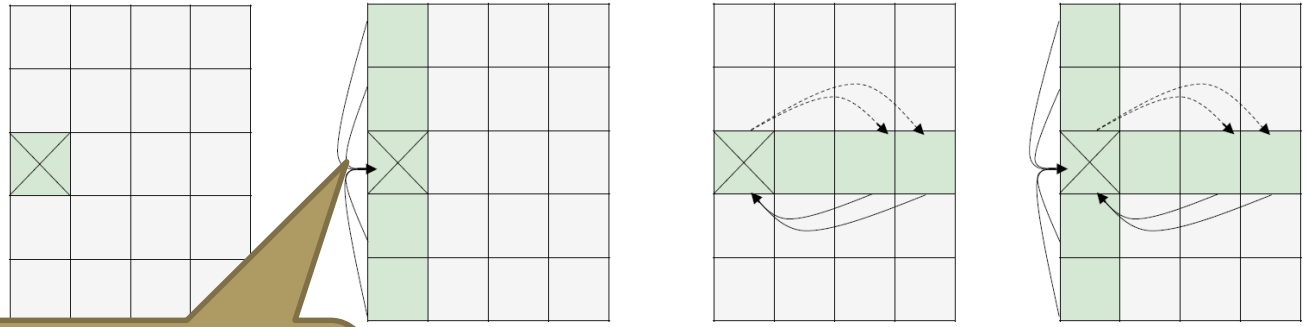


(a) Cell (b) Column (c) Row (d) Row-Column

Plain Cell Value
(various strategies)

Contexts and Annotation Modules

- 4 contexts used to create/filter annotations

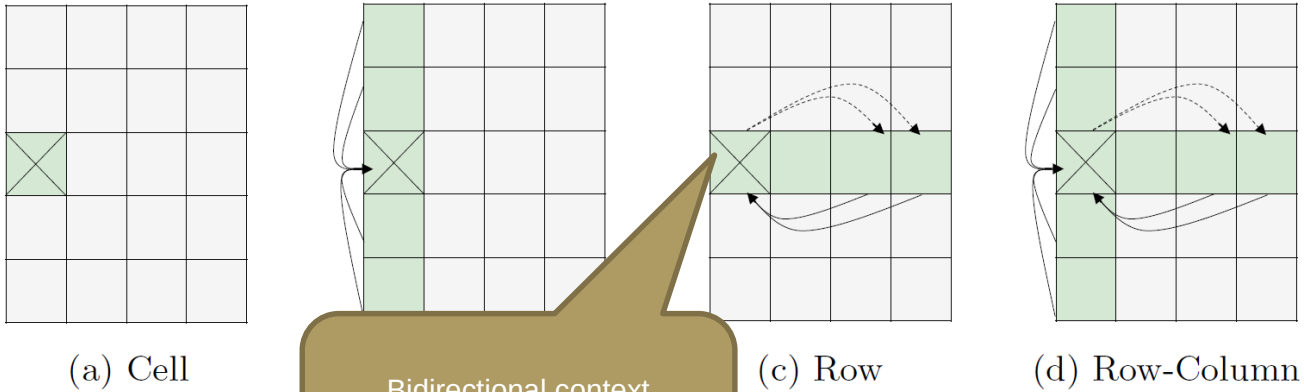


(b) Column (c) Row (d) Row-Column

Semantic Column Type
+ plain cell value
Assumes all cells in one column belong to the same type

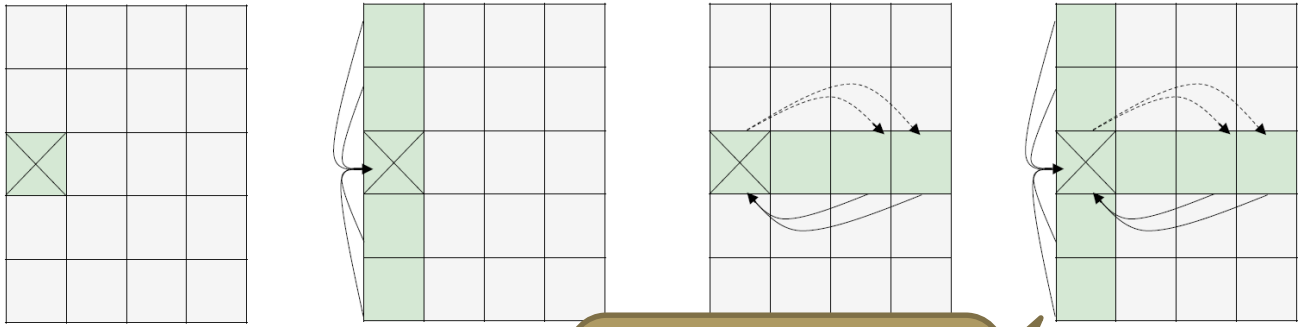
Contexts and Annotation Modules

- 4 contexts used to create/filter annotations



Contexts and Annotation Modules

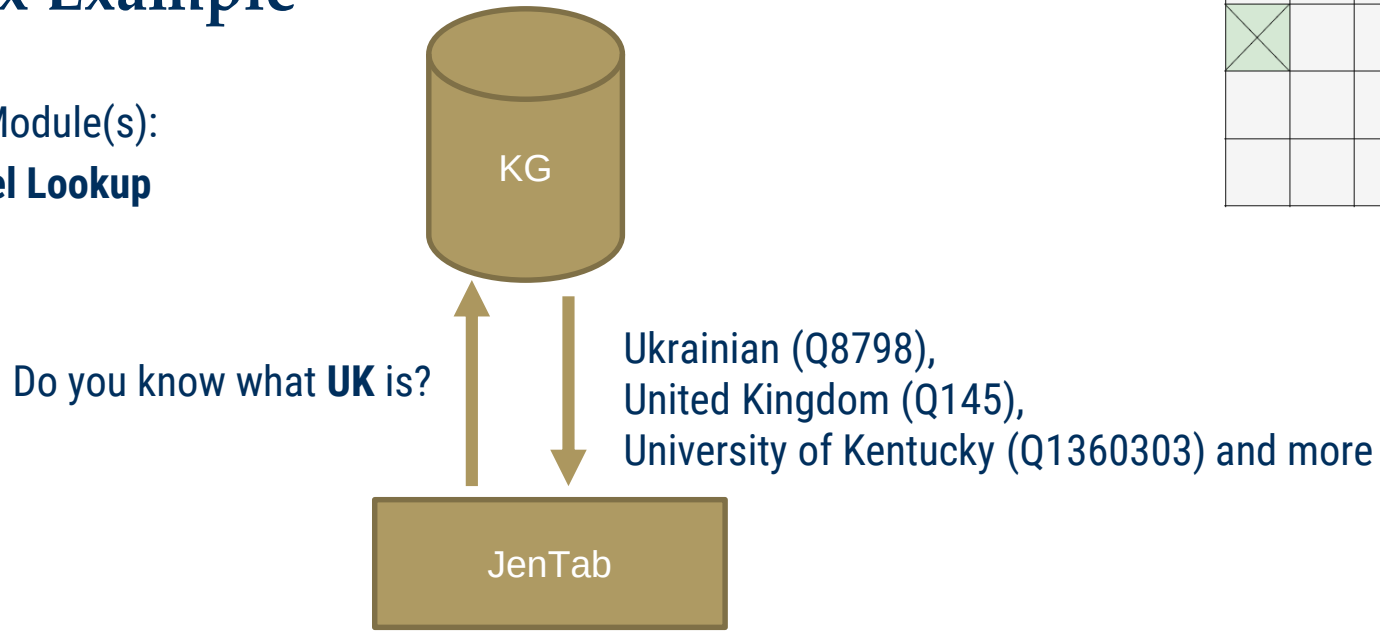
- 4 contexts used to create/filter annotations



Black Box Example

Annotation Module(s):

- **CEA Label Lookup**



* For complex cell values, e.g., 1st Global Opinion Leader's Summit, we try other strategies to create mappings. For example, look for each token in the cell as a standalone query.

CEA Label Lookup

- **4 strategies obtaining queries from cell value** + 2 handling spelling mistakes

Strategy	Priority	Method	Cell Example	Queries
Full Cell	1	Cleaned value as a query	Dainik Bhaskar	{Dainik Bhaskar}
Selective	2	parts before brackets	Mario's Super Picross (900 Wii Points)	{Mario's Super Picross}
Token	3	tokenize the cell values & exclude stopwords	Lost in Space	{Lost, Space}
All Token	3	tokenize the cell values, exclude stopwords & concatenate tokens ascending	Little House on the Prairie	{Little, House, Prairie, Little House, Little House Prairie, House Prairie}

CEA Label Lookup

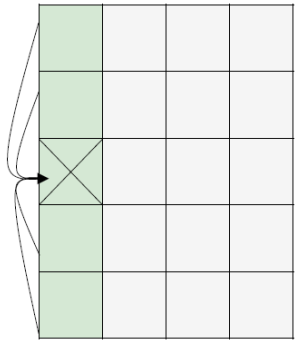
- 4 strategies obtaining queries from cell value + **2 handling spelling mistakes**

Generic Lookup	Autocorrect
Pre-computed	On demand
Executed before the actual pipeline	Invoked in cases of failure by Generic Lookup
Jaro-Winkler distance[1]	1-edit distance + word2vec ranking[2]
Highest priority (0)	Lowest priority (4)

[1] Winkler, W.E.: String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. (1990)

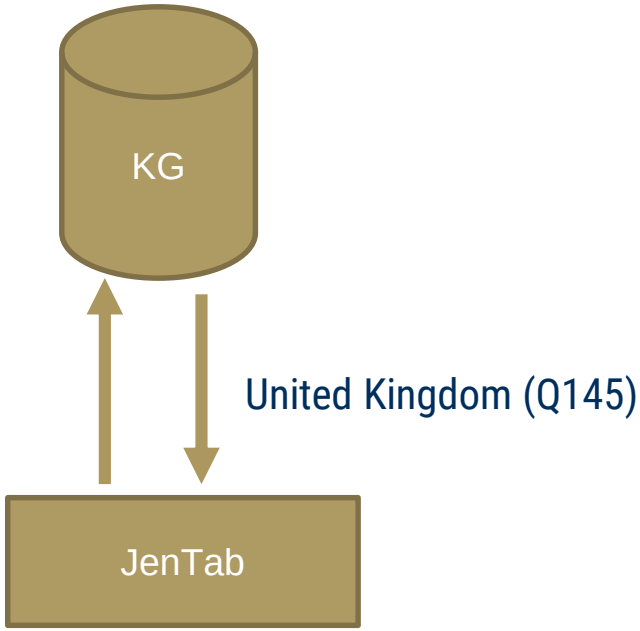
[2] https://www.kaggle.com/cpmpml/spell-checker-using-word2vec

Black Box Example



Annotation Module(s):

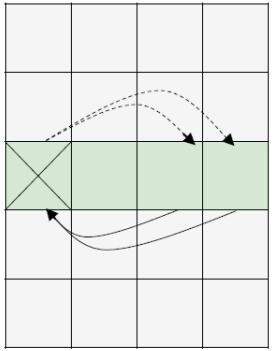
- **CEA by Column**



Do you know what **UK** is?
It is also a **country**

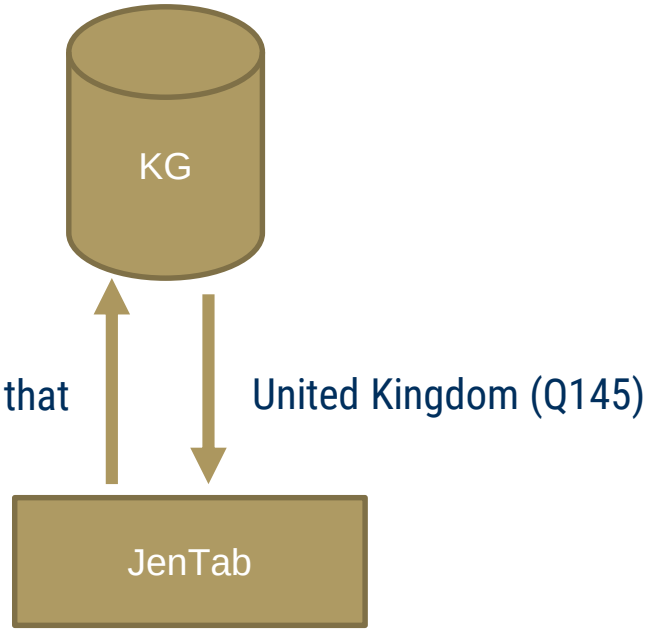
United Kingdom (Q145)

Black Box Example



Annotation Module(s):

- **CEA by row**
- **CEA by subject**



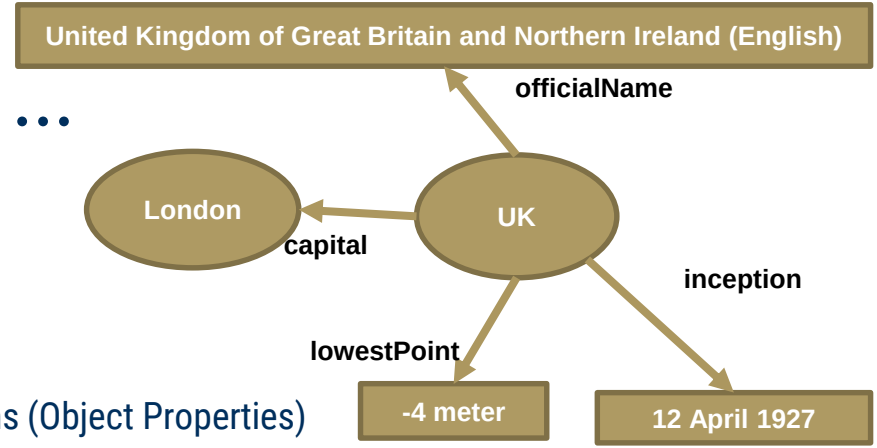
Do you know what is the **thing** that **has capital named London?**

United Kingdom (Q145)

Annotations Module Continued …

- **CTA**
  - Collects types for all retrieved cells annotations

Annotations Module Continued …



- **CTA**
  - Collects types for all retrieved cells annotations
- **CPA**
  - Collects properties for all retrieved cells annotations (Object Properties)
  - Fuzzy match properties with values only (Literal Properties)
    - **DATE**: try matching day, month and year parts only, ignore any other parts.
    - **QUANTITY**: support a margin of tolerance e.g., 10%
    - **STRING**: Calculates overlap between KG value and Table value. Consider a match if overlap > threshold.

Annotation Modules … Filter

- **CTA support**
  - Column types < support by cell candidates
  - Affects CTA and CEA candidates
- **CEA by unmatched properties**
  - Cell candidates have no matched properties
- **CEA by property support**
  - A generic form of the above
  - Considers support value
- **CEA by string distance**
  - Cell value vs. KG label value
  - Levenshtein distance > threshold

Annotation Modules … Select

- Picks the solution!

CEA

- **CEA by string similarity**
  - Selects the KG value with the closest Levenshtein distance
- **CEA by column**
  - Looks inside the same column for a similar value and pick its candidate

CPA

- **CPA by majority vote**
  - Picks most co-occurred property

CTA

- **CTA by LCS** vs. **CTA by majority**
- **CTA by direct parents**
- **CTA by popularity**

* LCS – Least Common Subsumer

Annotation Modules … Select

- Finds the solution!

CEA

- **CEA by string similarity**
  - Selects the KG value with the closet Levenshtein distance
- **CEA by column**
  - Looks inside the same column for a similar value and pick its candidate

CPA

- **CPA by majority vote**
  - Picks most co-occurred property

CTA

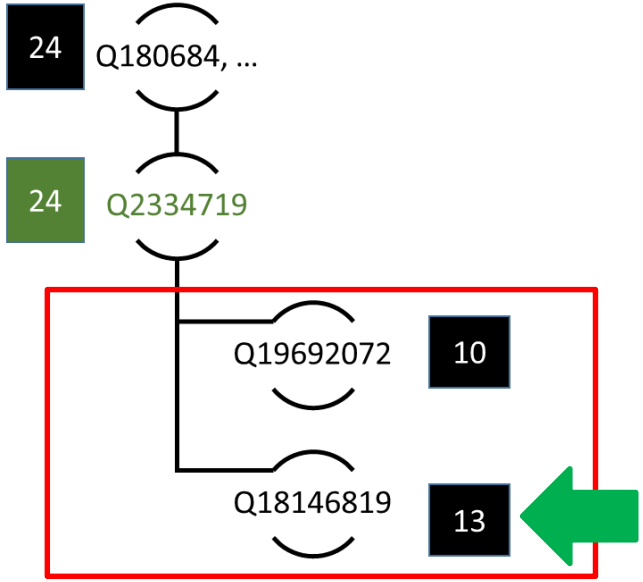
- **CTA by LCS** vs. **CTA by majority**
- **CTA by direct parents**
- **CTA by popularity**

* LCS – Least Common Subsumer

Select CTA

- **CTA by majority**

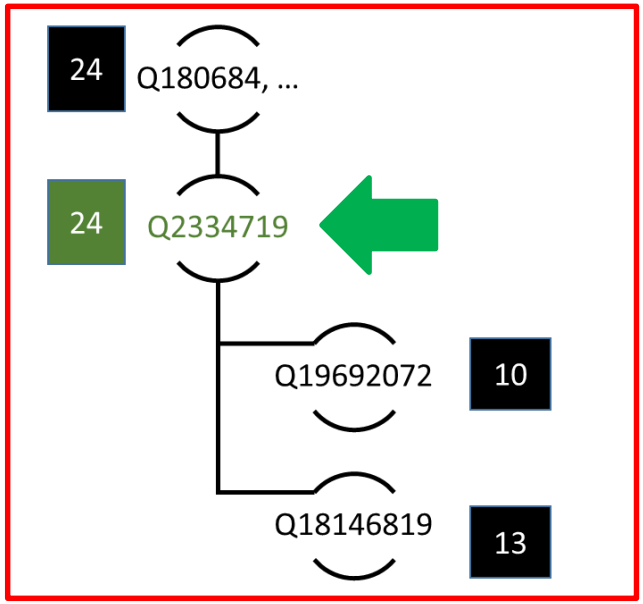
Col0	P31
Spaziano . Florida	Q18146819, Q19692072
Smith v/ Maryland	Q18146819, Q19692072
SEC v. Texas Gulf Sumphur Co.	Q2334719
Reieer v. Thompso	Q18146819, Q19692072
Reed v. Pennsylvania Railroad Compan|	Q18146819, Q2334719
Building Service Employees International Union Local 262 v/ Gazzam	Q18146819, Q19692072
Ramspeck v. Federal Trial Exainers Conference	Q18146819, Q2334719
Budk v. California	Q18146819, Q19692072
Cowma Dairy Company v. United States	Q18146819, Q2334719
Noswood v. Kirkpatrick	Q18146819, Q19692072
Mongomery Building & Construction Trades Council v. Ledbetter Erection Company	Q18146819, Q19692072
Southern Pacfic Company v. Gileo	Q18146819, Q19692072
Colgate-Palmolive-Peft Company v. National Labor Relations Board	Q18146819, Q19692072
Unitee States v. United States Smelting Refining	Q18146819, Q19692072
Poizzi v. Cowles Magazies	Q18146819, Q19692072



Select CTA

- **CTA by LCS**

Col0	P31
Spaziano . Florida	Q18146819, Q19692072
Smith v/ Maryland	Q18146819, Q19692072
SEC v. Texas Gulf Sumphur Co.	Q2334719
Reieer v. Thompso	Q18146819, Q19692072
Reed v. Pennsylvania Railroad Compan|	Q18146819, Q2334719
Building Service Employees International Union Local 262 v/ Gazzam	Q18146819, Q19692072
Ramspeck v. Federal Trial Exainers Conference	Q18146819, Q2334719
Budk v. California	Q18146819, Q19692072
Cowma Dairy Company v. United States	Q18146819, Q2334719
Noswood v. Kirkpatrick	Q18146819, Q19692072
Mongomery Building & Construction Trades Council v. Ledbetter Erection Company	Q18146819, Q19692072
Southern Pacfic Company v. Gileo	Q18146819, Q19692072
Colgate-Palmolive-Peft Company v. National Labor Relations Board	Q18146819, Q19692072
Unitee States v. United States Smelting Refining	Q18146819, Q19692072
Poizzi v. Cowles Magazies	Q18146819, Q19692072

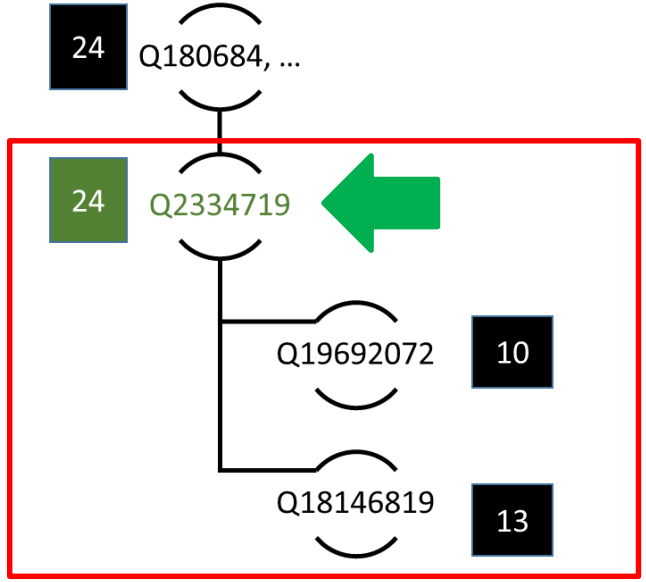


* LCS – Least Common Subsumer

Select CTA

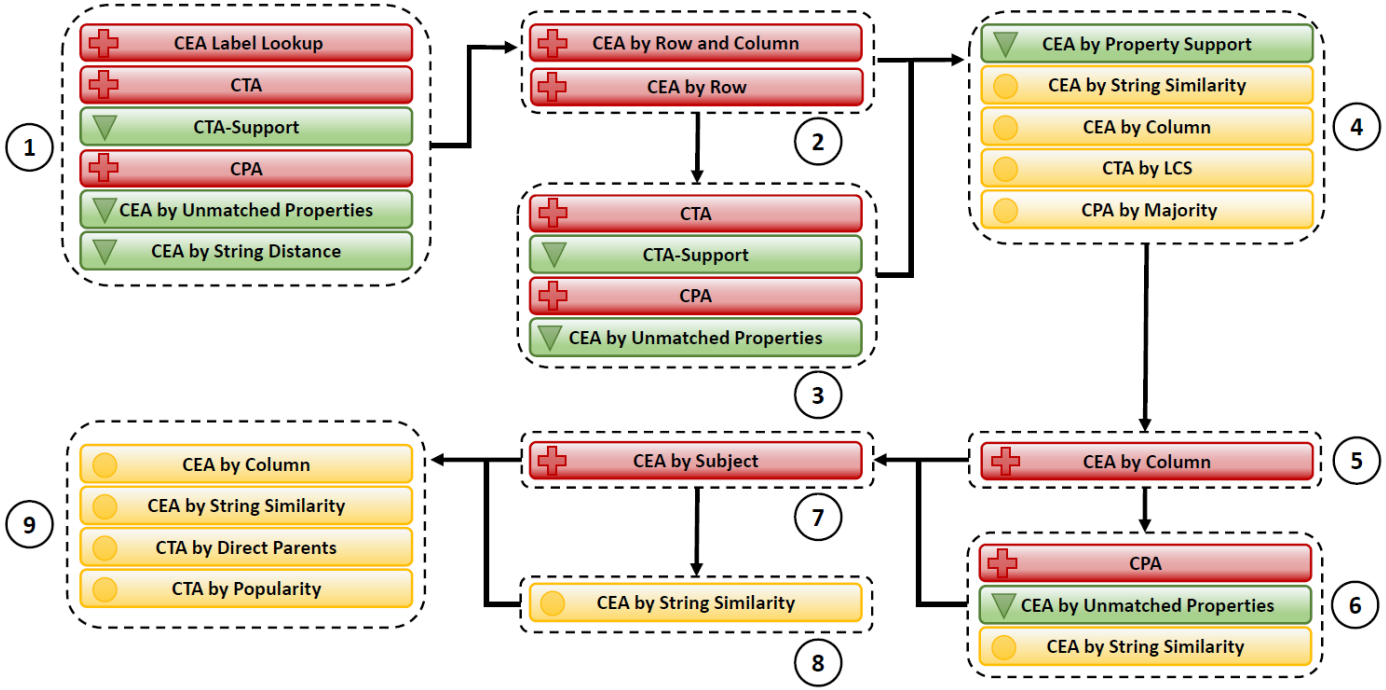
- **CTA by direct parents** (flatten leaves and parents + majority voting)

Col0	P31
Spaziano . Florida	Q18146819, Q19692072
Smith v/ Maryland	Q18146819, Q19692072
SEC v. Texas Gulf Sumphur Co.	Q2334719
Reieer v. Thompso	Q18146819, Q19692072
Reed v. Pennsylvania Railroad Compan|	Q18146819, Q2334719
Building Service Employees International Union Local 262 v/ Gazzam	Q18146819, Q19692072
Ramspeck v. Federal Trial Exainers Conference	Q18146819, Q2334719
Budk v. California	Q18146819, Q19692072
Cowma Dairy Company v. United States	Q18146819, Q2334719
Noswood v. Kirkpatrick	Q18146819, Q19692072
Mongomery Building & Construction Trades Council v. Ledbetter Erection Company	Q18146819, Q19692072
Southern Pacfic Company v. Gileo	Q18146819, Q19692072
Colgate-Palmolive-Peft Company v. National Labor Relations Board	Q18146819, Q19692072
Unitee States v. United States Smelting Refining	Q18146819, Q19692072
Poizzi v. Cowles Magazies	Q18146819, Q19692072



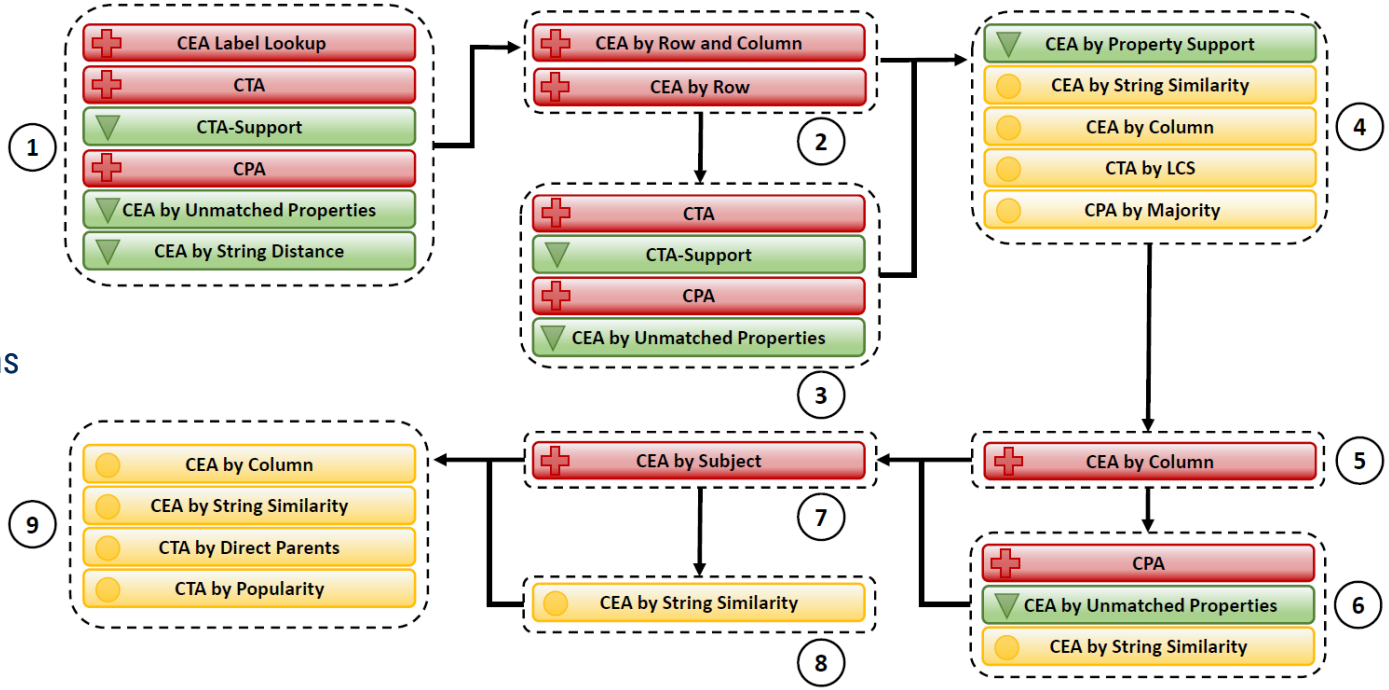
Sequence of Modules

- Create
  - Plus
  - Red
- Filter
  - Triangle
  - Green
- Select
  - Circle
  - Yellow



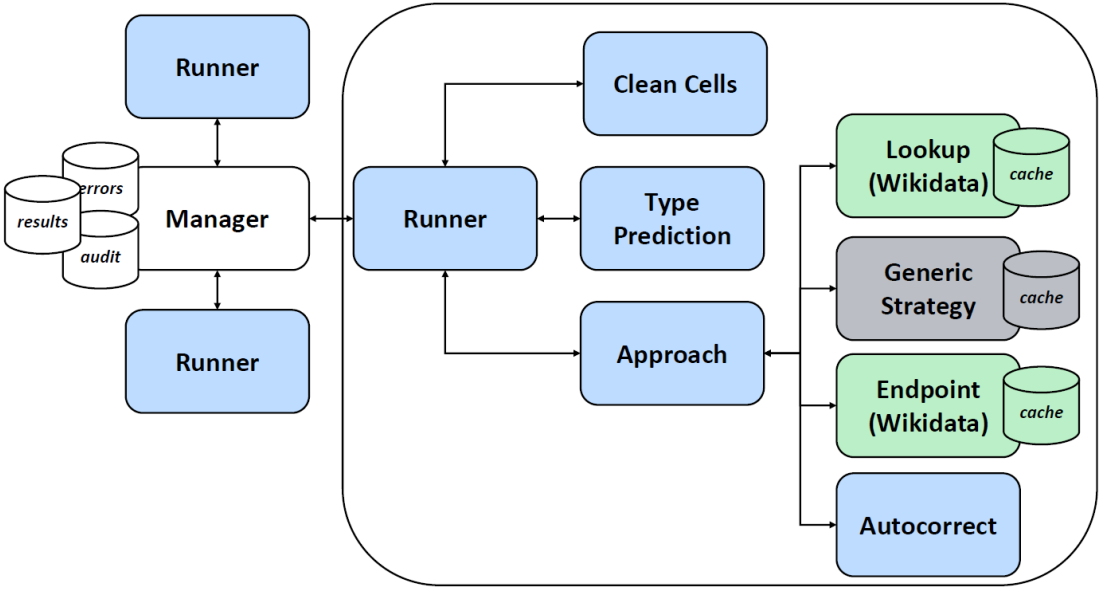
Sequence of Modules

- Group 1
  - Core Pipeline
  - No selection
- Group 9
  - Last resort
  - Backup solutions are invoked if failure of the previous methods



Architecture

- Distributed system
- Manager central point
- Isolated services
- Scalable
  - Fits large scale datasets
- Easily exchange
  - Data sources (KG substitution)
  - Approach



Datasets

- Automatically Generated
- Tough Tables
- 130K tables

Round	R1	R2	R3	R4
Tables #	34,294	12,173	62,614	22,390
Avg. Rows # (± Std Dev.)	7 ± 4	7 ± 7	7 ± 5	$109 \pm 11,120$
Avg. Columns # (± Std Dev.)	5 ± 1	5 ± 1	4 ± 1	4 ± 1
Avg. Cells # (± Std Dev.)	36 ± 20	36 ± 18	23 ± 18	$342 \pm 33,362$
Target Cells # (CEA)	985,110	283,446	768,324	1,662,164
Target Columns # (CTA)	34,294	26,726	97,585	32,461
Target Columns Pairs # (CPA)	135,774	43,753	166,633	56,475

Dataset Challenges

a) Missing metadata
b) Spelling mistakes
c) Ambiguity
d) Missing spaces
e) Inconsistent format
f) Nested pieces of information in Quantity fields
g) Redundant columns
h) Encoding issues
i) Noisy data
j) Missing values (nulls, empty strings and special characters)
k) Tables of excessive length

	Subject Column	← Object Columns / Properties →			
	Country	Inception (LITERAL)	Area (LITERAL)	Label (LITERAL)	Capital (IRI)
Raw Table	Country	Col1			a
	Egypt b	1922February, 28 d	1,010,407.87 km2 (... ft2)	Egypt	Cairo
	Germa?ny b	3 October 1990 (03.10.1990) e	357,400 km2 (... ft2) f	Germany g	TÃ¼bingen h i
	UK c	??	NA j	United Kingdom	London
	...	...	...	...	... k

Evaluation

- **CEA** & **CPA** metrics

Secondary Score

Primary Score

$$P = \frac{|correct\ annotations|}{|annotated\ cells|},\ R = \frac{|correct\ annotations|}{|target\ cells|},\ F1 = \frac{2 \times P \times R}{P + R}$$

- **CTA** metrics

$$cscore(\alpha) = \begin{cases} 1, & \text{if } \alpha \textit{ is in } \text{GT}, \\ 0.8^{d(\alpha)}, & \text{if } \alpha \textit{ is an ancestor of the } \text{GT}, \\ 0.7^{d(\alpha)}, & \text{if } \alpha \textit{ is a descendant of the } \text{GT}, \\ 0, & \text{otherwise} \end{cases}$$

Secondary Score

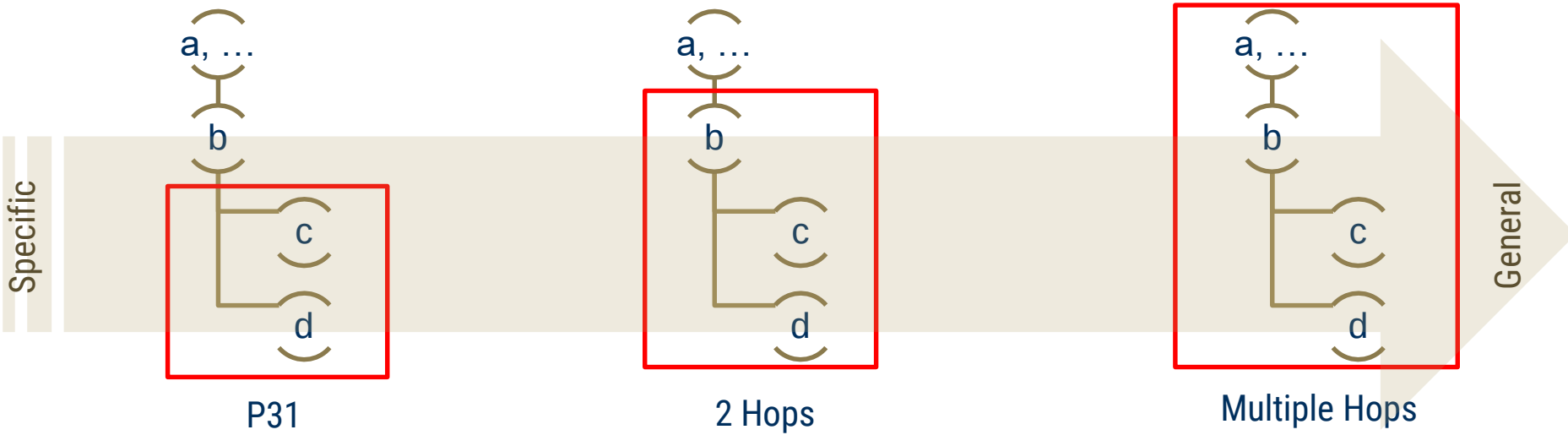
Primary Score

$$AP = \frac{\sum cscore(\alpha)}{|annotated\ cells|},\ AR = \frac{\sum cscore(\alpha)}{|target\ cells|},\ AF1 = \frac{2 \times AP \times AR}{AP + AR}$$

* GT is the ground truth
* AP Approximate Precision

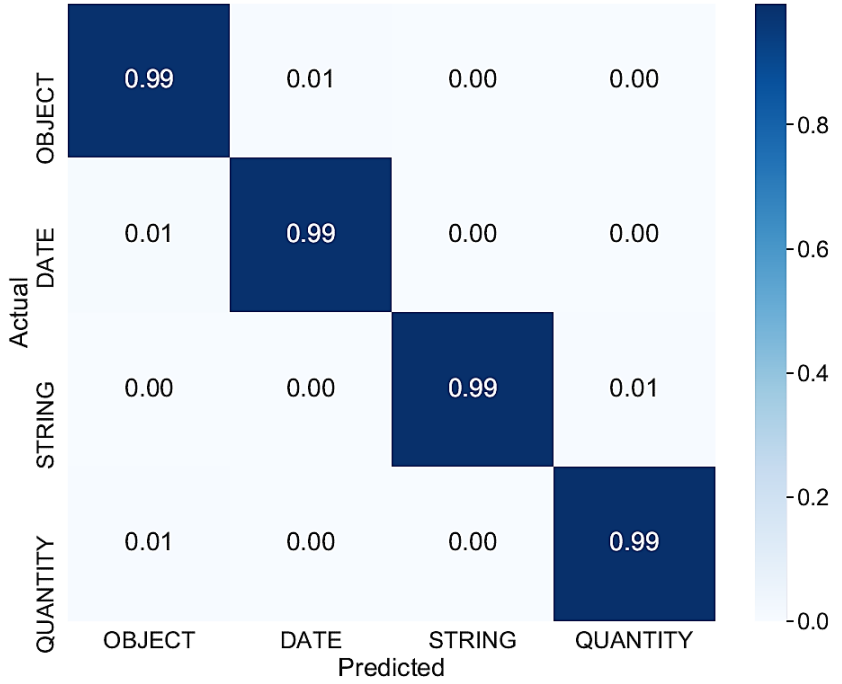
Experiments

- 3 Modes of CTA
  - Which one has the best scores?



Results

- Preprocessing
  - Type Prediction
  - Accuracy 99.0%



Results

- Generic Lookup
  - High coverage
  - Computationally expensive

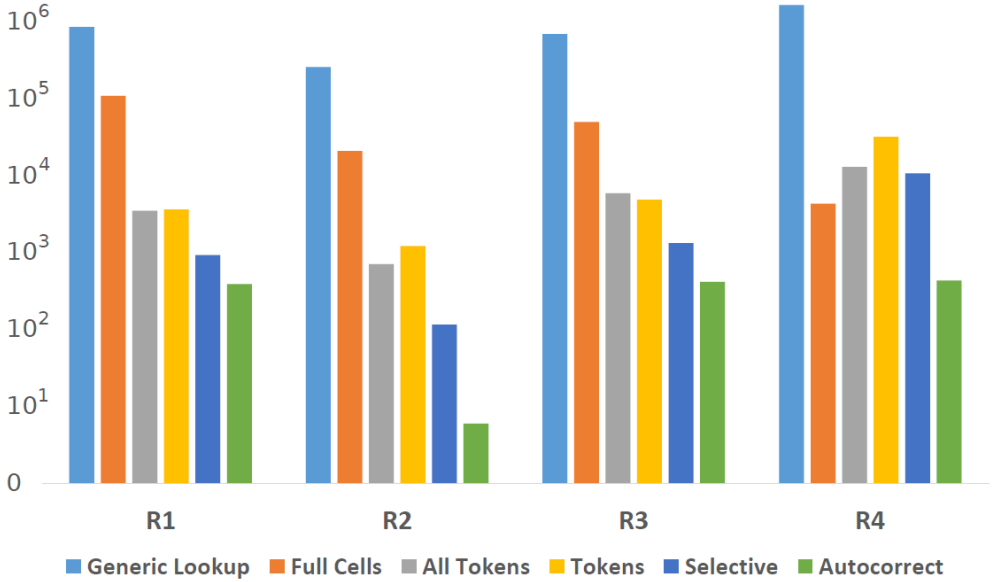
Rounds	Unique Labels	Matched (%)
R1	252,329	99.0
R2	132,948	98.9
R3	361,313	99.0
R4	533,015	96.8

https://github.com/fusion-jena/JenTab_precomputed_lookup

Results

- **Audit statistics for CEA**

- Reflects our priorities
- Various strategies capture a wide range of information inside cells

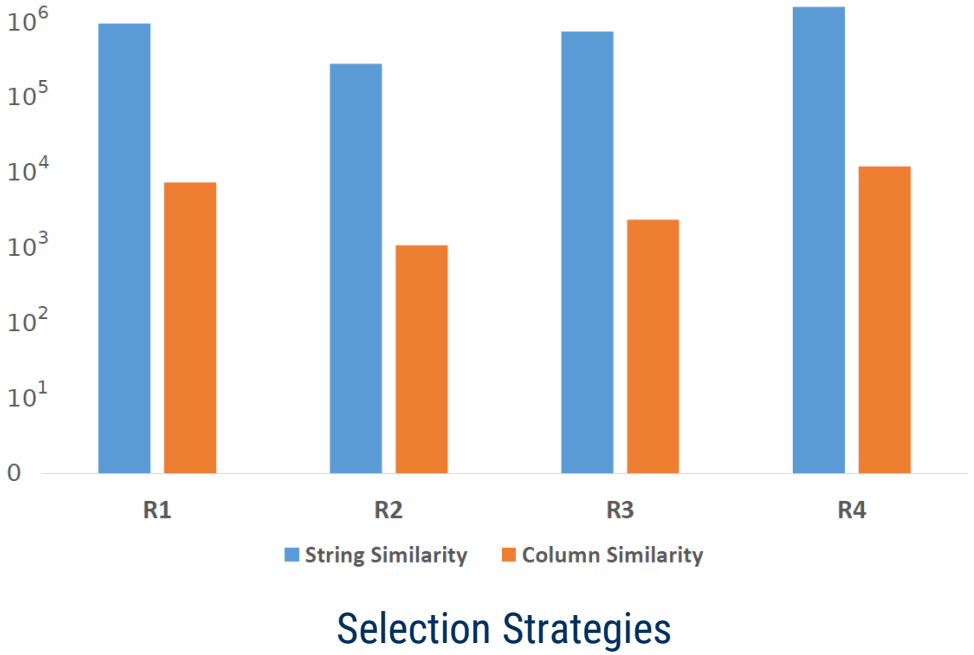


Creation Strategies

Results

- **Audit statistics for CEA**

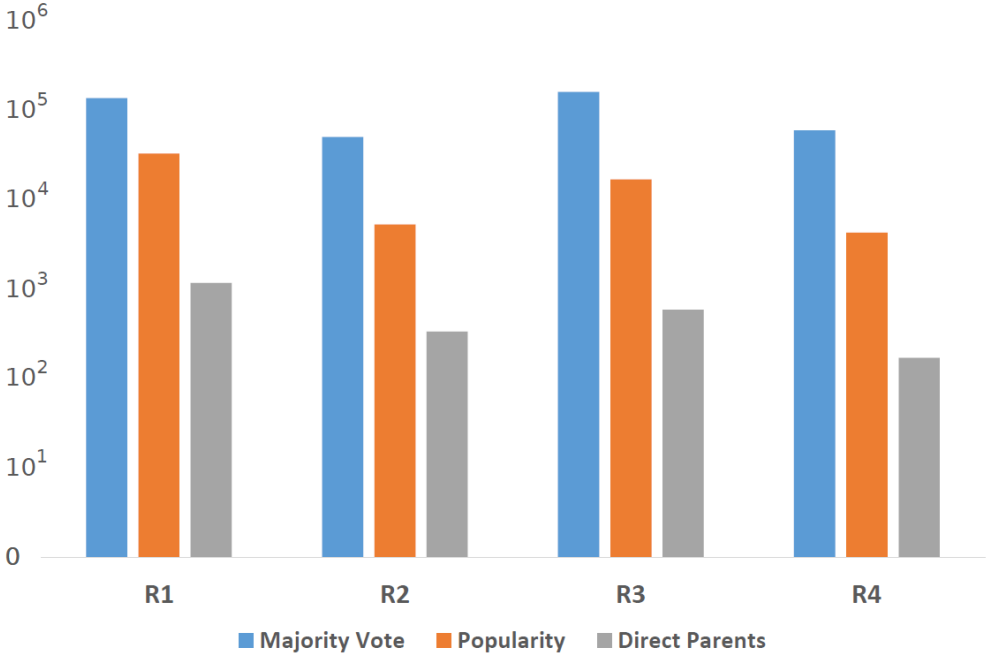
- String similarity is the dominant method
- Solves 38% more than column similarity
- The need of a backup method
- Some cells failed to have an annotation or annotation was removed by filter function



Results

- **Audit statistics for CTA**

- Majority vote is the dominant
- Backup solutions are frequently used

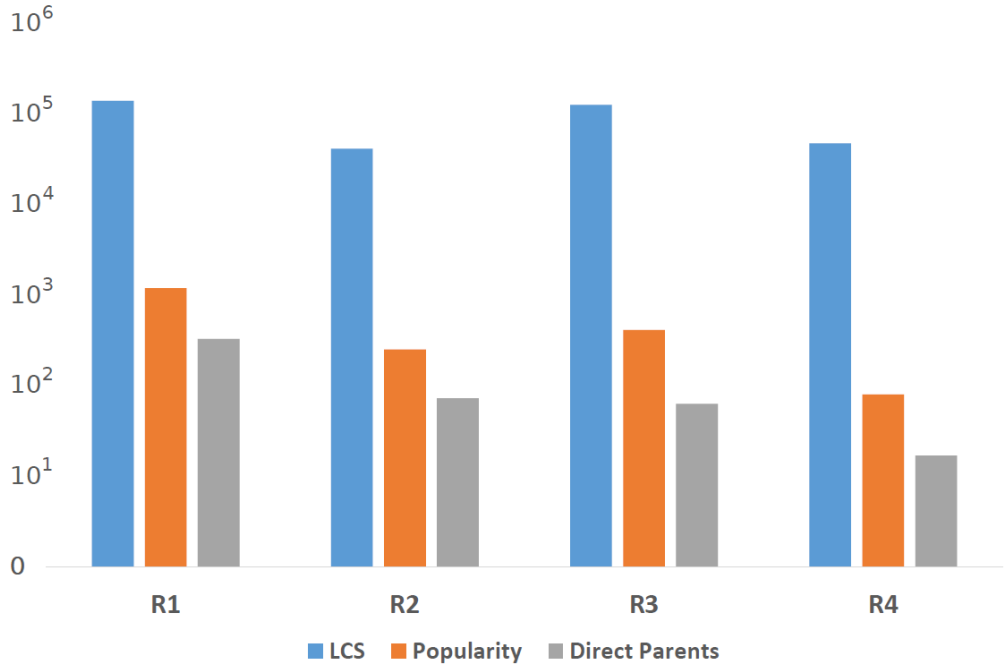


Mode: P31

Results

- **Audit statistics for CTA**

- LCS is the dominant
- Backup solutions are less frequently used
- LCS is more effective than Majority vote



Mode: 2 Hops

Results

- JenTab among the
  - Top 5 systems (CEA & CTA)
  - Top 3 systems (CPA)
- No Wikidata dump
- No generic search engines
  - SearX

	Automatically Generated Dataset						Tough Tables			
	CEA		CTA		CPA		CEA		CTA	
System	F1	Pr	F1	Pr	F1	Pr	F1	Pr	F1	Pr
JenTab (P31)	0.974	0.974	0.945	0.941	0.992	0.994	0.485	0.488	0.524	0.554
JenTab (2 Hops)	0.973	0.974	0.930	0.924	0.993	0.994	0.476	0.526	0.646	0.666
JenTab (Multiple Hops)	0.947	0.949	0.863	0.892	0.956	0.994	0.287	0.402	0.180	0.237
MTab4Wikidata	**0.993**	**0.993**	**0.981**	**0.982**	**0.997**	**0.997**	**0.907**	**0.907**	**0.728**	**0.730**
bbw	0.978	0.984	0.980	0.980	0.995	0.996	0.863	0.927	0.516	0.789
LinkingPark	0.985	0.985	0.953	0.953	0.985	0.986	0.810	0.811	0.686	0.687
DAGOBAH	0.984	0.985	0.972	0.972	0.995	0.995	0.412	0.749	0.718	0.747
SSL	0.833	0.833	0.946	0.946	0.924	0.924	0.198	0.198	0.624	0.669

Results

- JenTab among the
  - Top 5 systems (CEA & CTA)
  - Top 3 systems (CPA)
- No Wikidata dump
- No generic search engines
  - SearX
- Poor performance on 2T dataset
  - P31 is insufficient for hard cases

	Automatically Generated Dataset						Tough Tables			
	CEA		CTA		CPA		CEA		CTA	
System	F1	Pr	F1	Pr	F1	Pr	F1	Pr	F1	Pr
JenTab (P31)	0.974	0.974	0.945	0.941	0.992	0.994	0.485	0.488	0.524	0.554
JenTab (2 Hops)	0.973	0.974	0.930	0.924	0.993	0.994	0.476	0.526	0.646	0.666
JenTab (Multiple Hops)	0.947	0.949	0.863	0.892	0.956	0.994	0.287	0.402	0.180	0.237
MTab4Wikidata	**0.993**	**0.993**	**0.981**	**0.982**	**0.997**	**0.997**	**0.907**	**0.907**	**0.728**	**0.730**
bbw	0.978	0.984	0.980	0.980	0.995	0.996	0.863	0.927	0.516	0.789
LinkingPark	0.985	0.985	0.953	0.953	0.985	0.986	0.810	0.811	0.686	0.687
DAGOBAH	0.984	0.985	0.972	0.972	0.995	0.995	0.412	0.749	0.718	0.747
SSL	0.833	0.833	0.946	0.946	0.924	0.924	0.198	0.198	0.624	0.669

Results

- JenTab among the
  - Top 5 systems (CEA & CTA)
  - Top 3 systems (CPA)
- No Wikidata dump
- No generic search engines
  - SearX
- Multiple Hops
  - Too generic solutions
  - Lower scores

	Automatically Generated Dataset						Tough Tables			
	CEA		CTA		CPA		CEA		CTA	
System	F1	Pr	F1	Pr	F1	Pr	F1	Pr	F1	Pr
JenTab (P31)	0.974	0.974	0.945	0.941	0.992	0.994	0.485	0.488	0.524	0.554
JenTab (2 Hops)	0.973	0.974	0.930	0.924	0.993	0.994	0.476	0.526	0.646	0.666
JenTab (Multiple Hops)	0.947	0.949	0.863	0.892	0.956	0.994	0.287	0.402	0.180	0.237
MTab4Wikidata	**0.993**	**0.993**	**0.981**	**0.982**	**0.997**	**0.997**	**0.907**	**0.907**	**0.728**	**0.730**
bbw	0.978	0.984	0.980	0.980	0.995	0.996	0.863	0.927	0.516	0.789
LinkingPark	0.985	0.985	0.953	0.953	0.985	0.986	0.810	0.811	0.686	0.687
DAGOBAH	0.984	0.985	0.972	0.972	0.995	0.995	0.412	0.749	0.718	0.747
SSL	0.833	0.833	0.946	0.946	0.924	0.924	0.198	0.198	0.624	0.669

Results

- Execution Time
  - Time scoped
  - Faster convergence
  - R4 50% reduction

	R1		R2		R3		R4	
Mode	Days	Runners	Days	Runners	Days	Runners	Days	Runners
P31	0.5	4	2.5	4	1.5	6	2	4
2 Hops	1	4	1.2	4	2	4	1.1	8
Multi Hops	1	4	1.5	4	2.5	6	3.5	6

Conclusions

- JenTab toolkit*
  - Publicly available KG data sources
  - CFS pattern
  - 3 experiments of CTA
  - Detailed analysis of the 3 modes

* https://github.com/fusion-jena/JenTab

Future Work

- Optimize certain components that take substantial resources
  - Generic lookup
  - SPARQL queries
- Dig deeper into Tough Table dataset

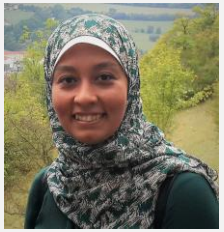
Acknowledgement

Dr. Kobkaew Opasjumruskit
Dr. Sheeba Samuel
Franziska Zander

Prof. Dr. Birgitta König-Ries
Prof. Dr. Joachim Denzler

Supported by the Carl Zeiss Foundation

Thank You!



Nora Abdelmageed

nora.abdelmageed@uni-jena.de

@NoraYoussef

Sirko Schindler

sirko.schindler@uni-jena.de