

JenTab Meets SemTab 2021's New Challenges

Nora Abdelmageed, Sirko Schindler

Friedrich Schiller University Jena, Germany
ISWC 2021

Semantic Table Annotation Tasks

Country	Area	Capital
Egypt	1,010,408	Cairo
Germany	357,386	Berlin

<https://www.wikidata.org/wiki/Q79>
<https://www.wikidata.org/wiki/Q183>

(a) CEA

Country	Area	Capital
Egypt	1,010,408	Cairo
Germany	357,386	Berlin

<https://www.wikidata.org/wiki/Q6256>

(b) CTA

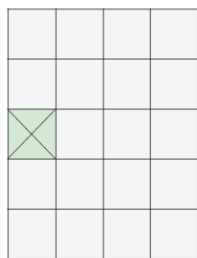
Country	Area	Capital
Egypt	1,010,408	Cairo
Germany	357,386	Berlin

<https://www.wikidata.org/wiki/Q5119>

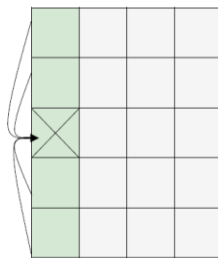
(c) CPA

Background

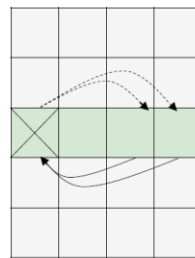
- JenTab uses **only** the public available **lookup API** and **SPARQL query endpoints** of the target knowledge graph
- Follows (CFS Pattern):
 - Create as many as candidates per task
 - Filter these candidates by using semantic context
 - Select the most appropriate candidate from the remaining
- Uses 4 contexts:



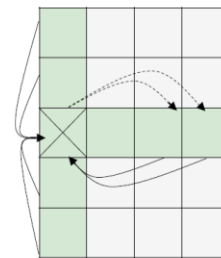
(a) Cell



(b) Column



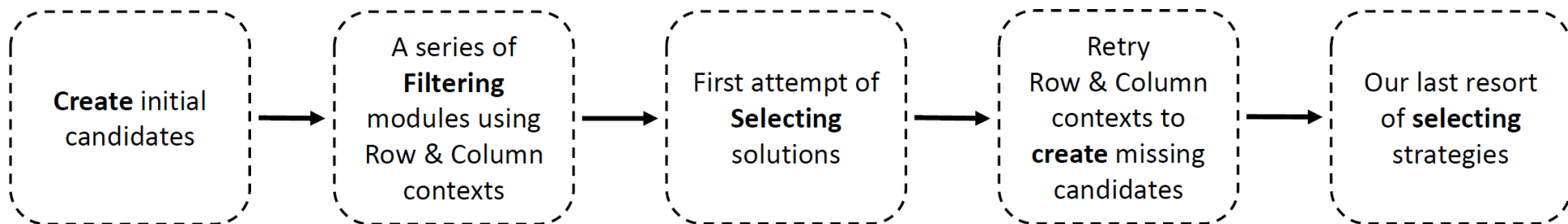
(c) Row



(d) Row-Column

Background – Default Pipeline

- Abstract View of **pipeline_full**



Outlook

1. Semantic Table Annotation Tasks
2. Background
3. New Design
 - System Architecture
 - Various Pipelines
4. Evolution
5. Experiences & Results
6. Conclusions & Future work

-
- ```

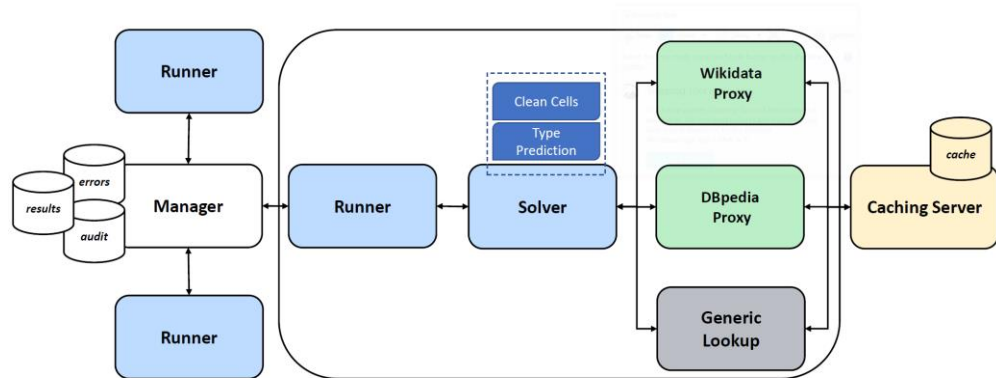
graph LR
 Manager[Manager] <--> Runner1[Runner]
 Manager <--> Runner2[Runner]
 Manager <--> Runner3[Runner]
 Manager --- results[(results)]
 Manager --- errors[(errors)]
 Manager --- audit[(audit)]
 Runner1 <--> Runner2
 Runner2 <--> Solver[Solver]
 Solver --- CleanCells[Clean Cells]
 Solver --- TypePrediction[Type Prediction]
 Solver <--> WikidataProxy[Wikidata Proxy]
 Solver <--> DBpediaProxy[DBpedia Proxy]
 Solver <--> GenericLookup[Generic Lookup]
 WikidataProxy <--> CachingServer[Caching Server]
 DBpediaProxy <--> CachingServer
 GenericLookup <--> CachingServer
 CachingServer --- cache[(cache)]

```

FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA Fakultät für Mathematik  
und Informatik

# Highlighting New Changes

- **Solver + Clean Cells + Type Prediction**
  - Less data transfers
  - Faster execution
- **DBpedia Proxy**
  - New requirement by SemTab 2021
- **Caching Server**
  - Reduced the redundancy made by individual caches
- **Generic Lookup**
  - Optimized, it is our primary auto-correction strategy



# Various Pipelines

- **pipeline\_essential**
  - Core parts of the **pipeline\_full** (no re-executions, faster)
- **pipeline\_no\_cpa**
  - Omits the CPA components (Create, Filter and Select)
- **pipeline\_keyTables**
  - Key-value pair grouping of the dataset
  - Key → clean table, without artificial noise
  - Value → noisy tables derived from that clean
  - Keys are solved by **pipeline\_full**
  - Values obtain their solutions via broadcasting

|                 |
|-----------------|
| Zooey Deschanel |
| Sarah McLachlan |

Clean table

|                   |
|-------------------|
| Zooey Deschanelle |
| Zooey Dechanelle  |
| Sarah McLaughlin  |
| Zooey Deschamel   |
| Zooey Dishonal    |
| Sarah McLauhclin  |
| Sarah Mcclaucklen |

Noisy table



# Various Pipelines

- **pipeline\_numeric**
  - Designed for tables with single object column (the subject) with a rest of numerical columns
  - Emphasizes the Row Context
    - Highest priority goes to candidates with the most support by their row
- **pipeline\_conditional**
  - Combines **pipeline\_numeric** and **pipeline\_full**
  - If the table structure meets (like above)
    - Solve using **pipeline\_numeric**
    - If coverage  $\geq 80\%$ , return solutions
  - Otherwise, Solve using **pipeline\_full**

# JenTab Evolution – SemTab 2021 Round 1

- Givens:
  - 180 tables, KG: Wikidata, Tasks: CEA & CTA
  - 180 tables, KG: DBpedia, Tasks: CEA & CTA
- Configuration:
  - DBpedia Proxy  $\rightarrow$  Lookup = Lookup<sup>[1]</sup> + spotlight<sup>[2]</sup>
  - **pipeline\_no\_cpa**  $\rightarrow$  No CPA required
  - **pipeline\_keyTables**  $\rightarrow$  Massive amount of noise, manual clean up and grouping data
- Findings:
  - DBpedia spotlight  $\rightarrow$  False positives
  - CTA might not be enough semantic context to filter candidates
    - CPA increases scores, precision.

[1] <http://lookup.dbpedia.org/api/search/PrefixSearch>

[2] <https://api.dbpedia-spotlight.org/en/candidates>

# JenTab Evolution – SemTab 2021 Round 2

- Givens:
  - HardTables, 1750 tables, KG: Wikidata, Tasks: CEA, CTA, CPA
  - BioTables, 110 tables, KG: Wikidata, Tasks: CEA, CTA, CPA
- Configuration :
  - **pipeline\_essential** → BioTables, wide tables with long strings, timeouts.
  - **pipeline\_full** → HardTables, ambiguity.
- Findings:
  - **pipeline\_full** is optimized and managed to run over BioTables
  - Best Results for both datasets

# JenTab Evolution – SemTab 2021 Round 3

- Givens:
  - HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  - BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  - GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA
- Configuration :
  - `pipeline_numeric` & `pipeline_conditional`
- Findings:
  - Most of the given tables follows
    - `Subject_col` & `[num1, num2, num3]`

# JenTab Evolution – SemTab 2021 Round 3

- Givens:
  - HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  - BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  - GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA
- Configuration:
  - **pipeline\_no\_cpa** → We couldn't relate the given object column to properties.
  - Dictionary based lookup for taxons, to solve abbreviations
    - Canna glauca → (C.glauca , ca.glauca)
  - Split nested entities and solve the first part only
    - David Eichenberg (University of Halle-Wittenberg) → **David Eichenberg** (University of Halle-Wittenberg)
- Findings:
  - Our nested entity solution is not optimal
    - The first entity may not be the target entity
  - Abbreviated taxons are ambiguous
    - C.glauca matches 48 plant species in Wikidata 14 Oct 2021



# JenTab Evolution – SemTab 2021 Round 3

- Givens:
  - HardTables, 7207 tables. KG: Wikidata. Tasks: CEA, CTA, CPA
  - BiodivTab, 50 tables. KG: Wikidata. Tasks: CEA, CTA
  - GitTables, 1101 tables. KG: DBpedia & schema.org. Tasks: CTA
- Configuration:
  - GitTables CTA targets structure mixes between column types (CTA) and properties (CPA)
    - We have split that into normal CTA and CPA targets
  - CTA/CPA depends on CEA
    - For our generated CTA, CEA is created.
  - Ontology is changed from dbpedia.org to schema.org in DBpedia Proxy
  - **pipeline\_essential**.
- Findings:
  - **pipeline\_full** suffers from timeout here
  - DBpedia Proxy managed to retrieve schema.org types but failed for properties.

# Experiences and Results

- SemTab2021 datasets. KGs: DBpedia (DPB), Wikidata (WD), and schema.org (SCH).

|                                | R1                |                   | R2               |                  | R3               |                        |                  |
|--------------------------------|-------------------|-------------------|------------------|------------------|------------------|------------------------|------------------|
|                                | 2T<br>DBP         | 2T<br>WD          | BioTables<br>WD  | HardTables<br>WD | BiodivTab<br>WD  | GitTables<br>DBP & SCH | HardTables<br>WD |
| Tables                         | 180               | 180               | 110              | 1,750            | 50               | 1,101                  | 7,207            |
| Avg. Rows # ( $\pm$ Std Dev.)  | 1,080 $\pm$ 2,798 | 1,080 $\pm$ 2,798 | 2,448 $\pm$ 193  | 17 $\pm$ 8       | 259 $\pm$ 743    | 58 $\pm$ 95            | 8 $\pm$ 5        |
| Avg. Cols # ( $\pm$ Std Dev.)  | 5 $\pm$ 2         | 4 $\pm$ 2         | 6 $\pm$ 1        | 3 $\pm$ 1        | 24 $\pm$ 13      | 16 $\pm$ 12            | 2 $\pm$ 1        |
| Avg. Cells # ( $\pm$ Std Dev.) | 4125 $\pm$ 10947  | 3952 $\pm$ 10129  | 14605 $\pm$ 2338 | 55 $\pm$ 32      | 4589 $\pm$ 10862 | 690 $\pm$ 1159         | 20 $\pm$ 15      |
| CEA #                          | 663,655           | 636,185           | 1,391,324        | 47,439           | 31,467           | 75,161* & 23,915*      | 58,948           |
| CTA #                          | 539               | 535               | 656              | 2,190            | 613              | 2,516 & 720            | 7,206            |
| CPA #                          | 359*              | 355*              | 546              | 3,835            | NA*              | 1,687* & 444*          | 10,694           |

## Experiences and Results

- Generic Lookup: Unique labels and ratio of resolved labels per round.

| Rounds | Dataset    | Target   | Unique Labels | Unmatched | Matched | Matched (%) |
|--------|------------|----------|---------------|-----------|---------|-------------|
| R1     | 2T         | Wikidata | 69,980        | 7,072     | 62,908  | 89.89%      |
| R1     | 2T         | DBpedia  | 66,340        | 7,172     | 59,168  | 89.19%      |
| R2     | HardTables | Wikidata | 249,625       | 600       | 249,025 | 99.76%      |
| R3     | HardTables | Wikidata | 47,809        | 944       | 46,865  | 98.03%      |
| R3     | GitTables  | DBpedia  | 37,780        | 21,253    | 16,527  | 43.75%      |

# Experiences and Results

- Primary, secondary scores, and Ranks for JenTab.

| Rounds  | Dataset    | Target   | CEA   |       |                 | CTA   |       |                 | CPA   |       |                 |
|---------|------------|----------|-------|-------|-----------------|-------|-------|-----------------|-------|-------|-----------------|
|         |            |          | F1    | Pr    | Rank            | AF1   | APr   | Rank            | F1    | Pr    | Rank            |
| Round 1 | 2T         | DBpedia  | 0.607 | 0.669 | 3 <sup>rd</sup> | 0.460 | 0.468 | 1 <sup>st</sup> | NA    | NA    | NA              |
| Round 1 | 2T         | Wikidata | 0.457 | 0.520 | 3 <sup>rd</sup> | 0.697 | 0.697 | 2 <sup>nd</sup> | NA    | NA    | NA              |
| Round 2 | HardTables | Wikidata | 0.966 | 0.967 | 4 <sup>th</sup> | 0.914 | 0.917 | 4 <sup>th</sup> | 0.996 | 0.997 | 2 <sup>nd</sup> |
| Round 2 | BioTables  | Wikidata | 0.857 | 0.858 | 4 <sup>th</sup> | 0.835 | 0.843 | 5 <sup>th</sup> | 0.899 | 0.899 | 2 <sup>nd</sup> |

F1 - F1 Score, Pr - Precision, AF1 - Average F1 Score, and APr - Average Precision.

# Conclusions

- JenTab matches table components to Wikidata/DBpedia KGs
- Updates of JenTab as a participant of SemTab2021
  - System redesign
  - A variety of pipelines based on dataset characteristics
- Our code, generic lookup & solution files are publicly available\*

\* <https://github.com/fusion-jena/JenTab>

## Future Work

- Change binary decision keep/remove candidate to a scoring mechanism
- Address is the lack of targets for specific tasks
  - GitTables
- Continuously improve the performance
  - Timeouts received from the public endpoints

# Acknowledgement

Sarah Böning  
Muhammad Abbady

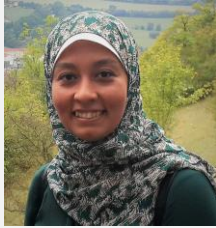
Prof. Dr. Birgitta König-Ries

The support is from the Carl Zeiss Foundation





# Thank You!



**Nora Abdelmageed**

[nora.abdelmageed@uni-jena.de](mailto:nora.abdelmageed@uni-jena.de)



@NoraYoussef

**Sirko Schindler**

[sirko.schindler@uni-jena.de](mailto:sirko.schindler@uni-jena.de)

