ISWC 2021

BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data

Nora Abdelmageed, Sirko Schindler, Birgitta König-Ries
Friedrich Schiller University Jena, Germany

ISWC 2021

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA Fakultät für Mathematik und Informatik

FUSION

MichaelStifelCenterJena for Data-Driven and Simulation Science

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

Motivation

- Semantic Table Annotation Tasks
  - SemTab 2019 – 2021
- General domain + Automatically Generated datasets
  - Except Tough Tables (manually curated)
- Benchmark from real-world data
- Domain-specific datasets have different characteristics
- Evaluating system on real-world data is needed

BiodivTab

- Domain-specific benchmark
- 50 tabular datasets of biodiversity research data
- Semantic Table Annotation Tasks
  - CEA and CTA tasks only
- Real Datasets + Data Augmentation
- Live Wikidata during September 2021

Data Collection

- Data sources (Biodiversity portals)

https://data.botanik.uni-halle.de/bef-china/
https://www.bexis.uni-jena.de/
https://data.world/

Data Exploration

- Biodiversity datasets have a unique set of challenges
  - Specimen Data
  - Numerical Fields
  - Nested Entities
    - `"David Eichenberg (University of Halle-Wittenberg)"`
  - Special Format
    - `"Species:abc Sub:xyz"`

Manual Annotation & Data Augmentation

- We have manually annotated Cells and column types for the collected datasets
  - 13 real from real data
  - In separate solution file for the quality assurance and multiple revisions
  - Manual disambiguation
  - Collect all matches remaining
- Partial revision by a Biodiversity expert
  - Fix annotations
- We have programmatically created new copies of the real data to increase the #tables and reduce manual annotation time
- Data augmentation is based on real data challenges
  - See Data Exploration

Format

- We follow the format of SemTab challenge datasets
  - **tables** → Collection of csv files to be annotated
  - **targets** → CEA, CTA and CPA (if exists) to be solved
  - **gt** (ground truth) → Solutions, hidden during the challenge

Final Assembly

- Anonymize table names using unique identifiers or uuid/python
- Combine the separate solution files into
  - **gt**
    - **cea**_BiodivTab_2021_gt.csv
    - **cta**_BiodivTab_2021_gt.csv
- Creates the "**targets**" version by dropping the solution column

SemTab 2021

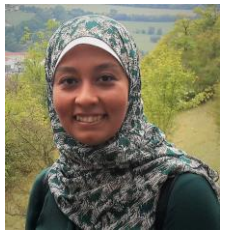
- SemTab 2021 organizers open the call for domain-specific datasets
- BiodivTab is accepted by SemTab organizers and made available for participants during Round 3, 2021.
- Avg scores*
  - CEA: 0.405
  - CTA: 0.228
- Hard dataset

* Announced in Ontology Matching (OM) Workshop, ISWC 2021

Acknowledgment

- Cornelia Fürstenau
- Andreas Ostrowski
- Samira Babalou

Thanks!



Nora Abdelmageed

nora.abdelmageed@uni-jena.de

@NoraYoussef

Birgitta König-Ries

Birgitta.Koenig-ries@uni-jena.de

@birgittaries

Sirko Schindler

sirko.schindler@uni-jena.de