BiodivTab: A Table Annotation Benchmark based on Biodiversity Research Data

Nora Abdelmageed[1−3][0000−0002−1405−6860] and Sirko Schindler[1,3][0000−0002−0964−4457] and Birgitta König-Ries[1−3][0000−0002−2382−9722] ⋆

[1] Heinz Nixdorf Chair for Distributed Information Systems
[2] Michael Stifel Center Jena, Germany
[3] Friedrich Schiller University Jena, Germany
nora.abdelmageed@uni-jena.de

Abstract. Semantic Table Annotation (STA) denotes the process of annotating a tabular dataset with concepts and relations from a given Knowledge Graph. The objective is to map individual table elements to their counterparts from the Knowledge Graph. The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) aims to establish a common framework for systems that tackle the process of STA. Since 2019, it has provided a set of benchmarks each year for evaluation. However, most of the provided datasets in the first two incarnations of the challenge are Automatically Generated (AG) and general domain datasets. This leaves the question open whether the developed systems can similarly be applied to real-world datasets that provide a different set of challenges. In this paper, we try to address this gap by introducing a domain-specific benchmark named BiodivTab. It consists of 50 datasets based on real-world biodiversity research data that have further been augmented. BiodivTab was made available to SemTab participants during Round 3 in the 2021 edition of the challenge.

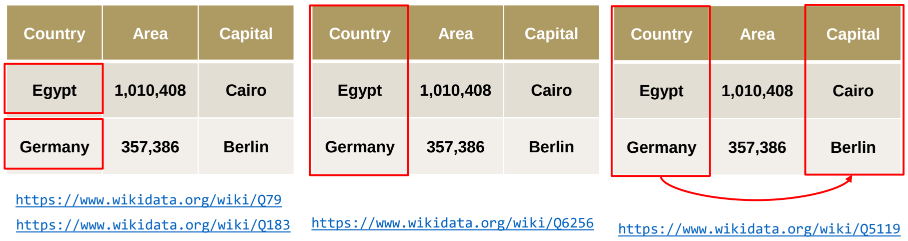
Keywords: Cell Entity Annotation · Column Type Annotation · Table Annotation · Benchmark · Biodiversity

1 Introduction

The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) has worked on establishing a community for Semantic Table Annotation (STA) tasks over the course of so far three editions: 2019 [9], 2020 [10], and 2021[4]. The challenge formulated three tasks for the STA that are illustrated by Figure 1. Each task matches a table component to its counterpart within a target Knowledge Graph (KG):

- Cell Entity Annotation (CEA) matches individual cells to entities.
- Column Type Annotation (CTA) assigns a semantic column type.
- Column Property Annotation (CPA) links column pairs using a semantic property.

[4] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/index.html



(a) CEA (b) CTA (c) CPA

Fig. 1: STA tasks as defined by SemTab (illustration from [1]).

The challenge establishes common standards for systems that tackle the problem of STA. Among the best-performing participants from the 2020 are MTab4Wikidata [17], LinkingPark [6], bbw [19], DAGOBAH [11], and JenTab [1].

The ultimate goal is systems that can annotate real-world datasets. However, the datasets introduced in the first two years of the challenge are Automatically Generated (AG) derived from different KGs [9, 10]. The ToughTables Dataset (2T) of 2020 is manually curated and focuses on the disambiguation of possible solutions [7]. The datasets employed, so far, adhere to no particular domain but represented a sample from a wide range of general-purpose data. On the other hand, domain-specific datasets pose specific challenges as witnessed, e.g., by evaluation campaigns in other domains like semantic web services evaluations [12]. So, ensuring that those challenges are covered, there is a demand for other domain-specific datasets based on real-world data. Furthermore, these benchmarks have to comply with the standards already in use by the community to easily highlight current shortcomings and encourage further efforts on these challenges.

In this paper, we introduce a domain-specific tabular benchmark named BiodivTab. We have collected real tables from the biodiversity domain and manually annotated them using the live edition of Wikidata [21] during September 2021, resulting in a human-level generated ground truth data for CEA and CTA tasks. Inspired by the challenges witnessed in the domain, we introduced artificially created variations to increase the number of included tables.

2 Benchmark Description

In this section, we explain the creation steps of BiodivTab, the data sources we have used, and the biodiversity-specific challenges encountered. Moreover, we describe the annotation phase, the data augmentation step, and the final assembly of the benchmark.

The included datasets were collected from three public repositories of biodiversity data: data.world[5], BEFChina[6] [5, 16, 20, 22], and BExIS[7] [4, 8, 13–15, 18]. We gathered

[5] https://data.world/

[6] https://data.botanik.uni-halle.de/bef-china/

[7] https://www.bexis.uni-jena.de/

a large collection of biodiversity-related datasets and verified their licenses[8] to ensure that they allow for the use in such a benchmark. Subsequently, we manually checked each of them concerning their suitability to the semantic table annotation tasks. In the process, we discarded datasets that predominantly contained, e.g., internal database "ID" columns, generic headers (e.g., "BEX_12"), or numerical columns without any further explanation or context. We consider those datasets next to impossible to annotate automatically and of little benefit to the community. Consequently, we decided to include only datasets containing a substantial amount of categorical information.

The datasets collected this way feature unique characteristics that can be summarized as follows:

- *Specimen Data*: The collected datasets contain observations of a particular specimen, e.g., a specific individual of a given species. This includes a multitude of properties of the specimen and their particular environment. The assembled data can only rarely be attributed to the general species.
- *Numerical Data*: Most of the collected datasets describe the specimen by various measurements in numerical form.
- *Abbreviations*: Species names may be given in an abbreviated format. For example, "Canna glauca", a particular kind of flower, is often referred to as "C.glauca" or "Ca.glauce".
- *Special Format*: Species names are identified by a combination of species and subspecies. For example, "species:Atrichum sub:subserratum" may be used instead of "Atrichum subserratum".

The identified challenges increase the difficulty during the semantic annotation process. Commonly, data is characterized by a single subject column (usually the leftmost one) with a group of other columns representing properties to the respective subject. Given that most of the data represent specimen data, those numerical/object fields do not necessarily relate to the general properties of species. As of the time of writing, Wikidata, the target KG of BiodivTab, contains no direct equivalent to the specimen data contained in the selected datasets. Thus, we could not annotate column-relations in the fashion of a CPA-task. As a consequence, BiodivTab does not include a CPA-task as of now. However, it might change in the future, if other KGs are supported, or Wikidata is extended accordingly. Instead, we provide ground truths only for CEA and CTA tasks. The special format found in the species names impedes the direct matching of cell values to labels of individual entities in the KG. One possible approach might be to handle these cases with a particular variety of misspellings. In fact, part of the cell values might be considered additional noise that has to be removed before matching.

After the data collection phase, we picked 13 tables to use in the annotation process. Our naming convention follows the schema of "dataSource_id", e.g., "befchina_1" represents the first dataset collected from BEFChina. The annotation itself is the most time-consuming part of the benchmark creation. To ensure the quality of mappings, we manually annotated the selected tables with entities assembled from Wikidata during September 2021, resulting in ground truth data for both CEA and CTA tasks. However, another set of annotations and checking the Inter-annotator Agreement [3] would be needed. Concerning CEA, we have marked possible candidate columns to annotate their cells. For each cell value, we assembled possible matches via Wikidata's built-in search. If more than one match has been found, we manually selected the most suitable

[8] https://github.com/fusion-jena/BiodivTab/blob/main/Read_Datasets_Licenses.md

ones to disambiguate the cell's semantics. If this still leaves more than one candidate, we keep them all and consider them true matches. Consequently, the provided ground truth contains all possible candidates that different systems could generate. We followed the same procedure for CTA. We maintain separate ground truth files to ease manual inspection, revision, and quality assurance for each table. E.g., "befchina_1" is annotated by two such files: "befchina_1_CEA" and "befchina_1_CTA". Biodiversity experts have partially revised these annotations. The structure of the ground truth files follows the format of SemTab. In particular, the solution files for CEA use a format of *filename, column id, row id, and ground truth*, whereas the ones for CTA employ a structure of *filename, column id, and ground truth.*

To increase the number of tables in our benchmark and reduce the human effort needed, we further resorted to data augmentation. It is a technique to increase the amount of data by adding slightly modified copies of already existing entries. In our context, we introduced challenges to the existing dataset based on our findings during the data collection and analysis phase (see the first paragraph). Since abbreviations are a common issue in biodiversity datasets, a column containing full species names was thus abbreviated accordingly. This strategy allowed us to (i) increase the number of included tables to 50 (almost 4× the number of just real tables), (ii) reduce the required human effort during annotation, and (iii) produce a benchmark that relies on real challenges instead of artificial ones.

The benchmark dataset consists of a set with 13 real tables and 37 augmented ones. We anonymized the file names of tables to use unique identifiers using Python's `uuid` functionalities in the process. Subsequently, we aggregated the individual solutions of CEA and CTA into one file per task resulting in *CEA_biodivtab_2021_gt.csv* and *CTA_biodivtab_2021_gt.csv*, respectively. From this, we generated corresponding "target-files" by removing the ground truth columns from these solution files. The benchmark tables and targets can now be published during the challenge for participants to solve. This follows the general approach of SemTab that hides the ground truth of STA tasks from participants during the challenge.

In 2021, the organizers published a call for domain-specific benchmarks to be used in that year's challenge. BiodivTab was submitted and accepted as one of these benchmarks. As a result, BiodivTab represented one of three benchmarks posed during the third round of 2021's SemTab challenge.

3 Conclusions & Future Work

We have introduced a tabular benchmark derived from biodiversity research data named BiodivTab. It consists of a collection of 50 tables. We have created BiodivTab by manually annotating 13 tables from real-world biodiversity datasets and adding 37 more tables by augmenting them with noise based on previously observed challenges. BiodivTab was submitted to and subsequently used in Round 3 of the 2021 SemTab challenge. Our benchmark is publicly available [2][9].

Future Work We see multiple directions to continue this work. We plan to include more biodiversity tables from other projects to cover a broader spectrum of the domain. In addition, ground truth data from other KGs, in particular domain-specific ones, can be provided.

[9] `https://github.com/fusion-jena/BiodivTab`

Acknowledgment

The authors thank the Carl Zeiss Foundation for the financial support of the project "A Virtual Werkstatt for Digitization in the Sciences (P5)" within the scope of the program line "Breakthroughs: Exploring Intelligent Systems" for "Digitization - explore the basics, use applications". We would like to especially thank our Biodiversity experts Cornelia Fürstenau and Andreas Ostrowski for feedback and validation of the created annotations. Last but not least, we would like to thank Samira Babalou for the fruitful discussions during the work.

The tables provided in this challenge are based on real-world biodiversity research datasets, but have been adapted for the challenge. In the form provided here, they may be used for the challenge, only. Any publication on challenge results needs to contain citations of the underlying datasets. The list of original datasets is available within our GitHub repository.

References

1. Abdelmageed, N., Schindler, S.: JenTab: Matching Tabular Data to Knowledge Graphs. In: SemTab@ ISWC. pp. 40–49 (2020)
2. Abdelmageed, N., Schindler, S., Knig-Ries, B.: fusion-jena/BiodivTab (Oct 2021). https://doi.org/10.5281/zenodo.5584180
3. Artstein, R.: Inter-annotator Agreement, pp. 297–313. Springer Netherlands (2017), `https://doi.org/10.1007/978-94-024-0881-2_11`
4. Boeddinghaus, R., Marhan, S., Berner, D., Boch, S., Fischer, M., Kattge, J., Klaus, V., Kleinebecker, T., Oelmann, Y., Prati, D., Schäfer, D., Schöning, I., Schrumpf, M., Sorkau, E., Kandeler, E., Manning, P., Kandeler, E.: Plant functional trait shifts explain concurrent changes in the structure and function of grassland soil microbial communities (2017). https://doi.org/10.25829/bexis.24867-1.1.23
5. Bruelheide, H., Eichenberg, D., Kröber, W., Böhnke, M., Ristok, C.: Main Experiment: Leaf traits and chemicals from individual trees in the Main Experiment (Site A & B) (2012), `https://china.befdata.biow.uni-leipzig.de/datasets/323`
6. Chen, S., Karaoglu, A., Negreanu, C., Ma, T., Yao, J.G., Williams, J., Gordon, A., Lin, C.Y.: LinkingPark: An Integrated Approach for Semantic Table Interpretation. In: SemTab@ ISWC. pp. 65–74 (2020)
7. Cutrona, V., Bianchi, F., Jimnez-Ruiz, E., Palmonari, M.: Tough Tables: Carefully Evaluating Entity Linking for Tabular Data (Nov 2020). https://doi.org/10.5281/zenodo.4246370
8. Fischer, M., Nauss, T., Tschapka, M., Weisser, W., Müller, J.: Aggregated species richness and habitat heterogeneity variables for testing the habitat-heterogeneity hypothesis, 2006-2018 (2020). https://doi.org/10.25829/bexis.25126-1
9. Hassanzadeh, O., Efthymiou, V., Chen, J., Jimnez-Ruiz, E., Srinivas, K.: SemTab2019: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching - 2019 Data Sets (Oct 2019). https://doi.org/10.5281/zenodo.3518539
10. Hassanzadeh, O., Efthymiou, V., Chen, J., Jimnez-Ruiz, E., Srinivas, K.: SemTab 2020: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Data Sets (Nov 2020). https://doi.org/10.5281/zenodo.4282879
11. Huynh, V.P., Liu, J., Chabot, Y., Labbé, T., Monnin, P., Troncy, R.: DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In: SemTab@ ISWC. pp. 27–39 (2020)

12. Küster, U., König-Ries, B.: Towards standard test collections for the empirical evaluation of semantic web service approaches. International Journal of Semantic Computing **2**(03), 381–402 (2008)
13. Leonhardt, S., Peters, B., Keller, A.: Amino acids in pollen of Osmia bicornis larval provisions 2017-2018 (2020). https://doi.org/10.25829/bexis.27228-4
14. Leonhardt, S., Peters, B., Keller, A.: Fatty acids in pollen of Osmia bicornis larval provisions 2017-2018 (2020). https://doi.org/10.25829/bexis.27227-2
15. Leonhardt, S., Peters, B., Keller, A.: Trap nesting solitary bee species measured on all grassland VIPs 2017-2018 (2020). https://doi.org/10.25829/bexis.27226-4
16. Nadrowski, K.: Deviations from stem breaking probabilities at species level (2013), `http://china.befdata.biow.uni-leipzig.de/datasets/327`
17. Nguyen, P., Yamada, I., Kertkeidkachorn, N., Ichise, R., Takeda, H.: MTab4Wikidata at SemTab 2020: Tabular Data Annotation with Wikidata. In: SemTab@ ISWC. pp. 86–95 (2020)
18. Seibold, S., Gošner, M., Simons, N., Blüthgen, N., Müller, J., Ambarli, D., Ammer, C., Bauhus, J., Fischer, M., Fürstenau, C., Habel, J.C., Linsenmair, K.E., Nauss, T., Ostrowski, A., Penone, C., Prati, D., Schall, P., Schulze, E.D., Vogt, J., Wöllauer, S., Weisser, W.: Arthropod data from 150 grassland plots, 2008-2017, and 140 forest plots, 2008-2016, used in "Arthropod decline in grasslands and forests is associated with drivers at landscape level", Nature (2019). https://doi.org/10.25829/bexis.25786-1.3.11
19. Shigapov, R., Zumstein, P., Kamlah, J., Oberländer, L., Mechnich, J., Schumm, I.: bbw: Matching csv to wikidata via meta-lookup. In: CEUR Workshop Proceedings. vol. 2775, pp. 17–26. RWTH (2020)
20. Staab, M., Schuldt, A., Assmann, T., Bruelheide, H., Klein, A.: Ant community structure during forest succession in a subtropical forest in South-East China pp. 32–40 (2014)
21. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (sep 2014). https://doi.org/10.1145/2629489
22. Wubet, T., Wu, Y., Buscot, F.: Soil Fungal metagenome from 12 CSPs based on the fungal ITS rDNA pyrotags (2013), `http://china.befdata.biow.uni-leipzig.de/datasets/397`