

BiodivTab: Semantic Table Annotation Benchmark Construction, Analysis, and New Additions

Nora Abdelmageed^{1,2,3}, Sirko Schindler^{1,3} and Birgitta König-Ries^{1,2,3}

¹Heinz Nixdorf Chair for Distributed Information Systems

²Michael Stifel Center Jena

³Friedrich Schiller University Jena, Jena, Germany

Abstract

Systems that annotate tabular data semantically have witnessed increasing attention from the community in recent years; this process is commonly known as Semantic Table Annotation (STA). Its objective is to map individual table elements to their counterparts from a Knowledge Graph (KG). Individual cells and columns are assigned to KG entities and classes to disambiguate their meaning. STA-systems achieve high scores on the existing, synthetic benchmarks but often struggle on real-world datasets. Thus, realistic evaluation benchmarks are needed to enable the advancement of the field. In this paper, we detail the construction pipeline of BiodivTab, the first benchmark based on real-world data from the biodiversity domain. In addition, we compare it with the existing benchmarks. Moreover, we highlight common data characteristics and challenges in the field. BiodivTab is publicly available¹ and has 50 tables as a mixture of real and augmented samples from biodiversity datasets. It has been applied during the SemTab 2021 challenge, and participants achieved F1-scores of at most $\sim 60\%$ across individual annotation tasks. Such results show that domain-specific benchmarks are more challenging for state-of-the-art systems than synthetic datasets.

Keywords

Benchmark, Tabular Data, Cell Entity Annotation, Column Type Annotation, Knowledge Graph Matching

1. Introduction

Systems that tackle annotating tabular data semantically have gained increasing attention from the community in recent years. Semantic Table Annotation (STA) tasks map individual table elements to their counterparts from a Knowledge Graph (KG) such as Wikidata [1], and DBpedia [2]. Here, individual cells and columns are assigned to KG entities and classes to disambiguate their meaning. The Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)² opened the call for semantic interpretation of tabular data inviting automated annotation systems. It established a common standard for evaluating those systems [3, 4, 5]. Most of its benchmarks are auto-generated with no particular domain focus [6, 7, 8, 9, 10, 11]. The ToughTables Dataset (2T) [12], introduced in the 2021 edition of the challenge, is the only exception involving manual curation but is still artificially derived from general domain data. Real-world and domain-specific datasets pose different challenges as witnessed,


¹<https://github.com/fusion-jena/BiodivTab>


Ontology Matching @ISWC 2022

✉ nora.abdelmageed@uni-jena.de (N. Abdelmageed); sirko.schindler@uni-jena.de (S. Schindler);

birgitta.koenig-ries@uni-jena.de (B. König-Ries)

ORCID [0000-0002-1405-6860](https://orcid.org/0000-0002-1405-6860) (N. Abdelmageed); [0000-0002-0964-4457](https://orcid.org/0000-0002-0964-4457) (S. Schindler); [0000-0002-2382-9722](https://orcid.org/0000-0002-2382-9722) (B. König-Ries)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

²<https://www.cs.ox.ac.uk/isg/challenges/sem-tab/>

e.g., by evaluation campaigns in other domains like semantic web services evaluations [13]. Therefore, the development of STA systems has to be accompanied by suitable benchmarks to make them applicable in real-world scenarios. Such benchmark should reflect idiosyncrasies and challenges immanent in different domains.

In this paper, we focus on one important domain: Biodiversity is the assortment of life on Earth covering evolutionary, ecological, biological, and social forms. It is imperative to monitor the current state of biodiversity and its change over time and understand the forces driving it to preserve life in all its varieties. The recent IPBES worldwide evaluation³ predicts a dramatic decrease in biodiversity, causing an obvious decay in vital ecological functions. An expanding volume of heterogeneous data, especially tables, is produced and publicly shared in the biodiversity domain. Tapping into this wealth of information requires two main steps: On the one hand, individual datasets have to be fit for (re)use – a requirement that resulted in the FAIR principles [14]. On the other hand, complex analyses often require data of different sources, e.g., to examine the various interdependencies among processes in an ecosystem. The involved datasets need to be integrated which requires a certain degree of harmonization and mappings between them [15]. The semantic annotation of the respective datasets can substantially support both goals.

Our unique contributions in this paper over our previous work [16] are as follows:

- Detailed explanation of the creation and data augmentation of BiodivTab.
- An extensive discussion of idiosyncrasies and challenges in biodiversity datasets.
- The creation of a new ground truth based on DBpedia.
- A characterization of BiodivTab including concepts covered.
- Evaluation of BiodivTab compared to other existing benchmarks.
- Applications of BiodivTab.

The remainder of this paper is organized as follows: Section 2 summarizes the required background. We detail the construction of BiodivTab in Section 3. Section 4 provides an evaluation of BiodivTab. Finally, we conclude in Section 5.

2. Background

Semantic Table Annotation: The SemTab challenge has provided a community forum for STA tasks over the course of so far four editions: 2019-2021 [6, 7, 17], and 2022⁴. The challenge established common standards to evaluate different approaches in the field. It captures increasing attention from the community. The best-performing participants in 2021 are DAGOBAN [18], MTab [19], and JenTab [20]. The challenge formulated three tasks illustrated by Figure 1. Each task matches a table component to its counterpart within a target 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.

Existing Benchmarks: The ultimate goal for STA-systems is to annotate real-world datasets. However, the datasets introduced in the first two years of the challenge are synthetic derived from different KGs [6, 7]. In 2020, the 2T dataset [12] is manually curated and focuses on the

³<https://ipbes.net/global-assessment>

⁴<https://sem-tab-challenge.github.io/2022/>

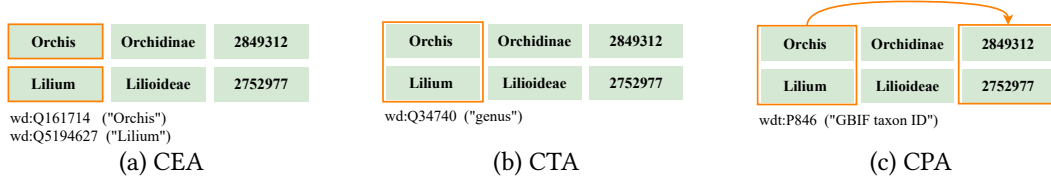


Figure 1: STA-tasks as defined by SemTab using a biodiversity example⁵.

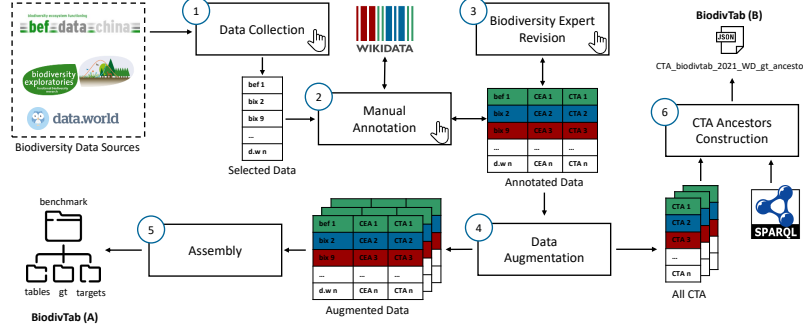


Figure 2: Steps of BiodivTab construction.

disambiguation of possible annotation solutions. 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 [13]. So, to ensure that those challenges are covered, there is a demand for domain-specific datasets based on real-world data. Such benchmarks have to comply with the standards already in use by the community to easily highlight current shortcomings and encourage further efforts on this topic.

3. BiodivTab Construction

In this section, we explain the creation of BiodivTab, and the data sources used. Moreover, we describe the manual annotation phase involving biodiversity experts, the data augmentation step, and the final assembly and release of the benchmark. Figure 2 summarizes the construction of BiodivTab, we detail in the following.

3.1. Data Collection

We decided on three data repositories that are well established for the ecological data: BExIS⁶, BEFChina⁷, and data.world⁸. We queried these portals using 20 keywords, e.g., abundance, and species, from our previous work [21]. Subsequently, we manually checked all of them regarding their suitability to the STA-tasks. We discarded datasets that contained a majority of,

⁵We use the following prefixes throughout this paper: dbr : <http://dbpedia.org/resource/>, dbo : <http://dbpedia.org/ontology/>, rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#>, rdfs : <http://www.w3.org/2000/01/rdf-schema#>, wd : <http://www.wikidata.org/entity/>, wdt : <http://www.wikidata.org/prop/direct/>, and owl : <http://www.w3.org/2002/07/owl#>

⁶<https://www.bexis.uni-jena.de/>

⁷<https://data.botanik.uni-halle.de/bef-china/>

⁸<https://data.world/>

Table 1

Prevalence of challenges among the selected datasets.

Dataset	Nested Entities	Acronyms	Typos	Numerical Data	Missing Values	Lack of Context	Synecdoche	Specimen Data
dataworld_1	○	●	○	●	●	●	○	●
dataworld_2	●	●	○	●	●	●	○	●
dataworld_4	●	●	●	●	●	●	●	●
dataworld_6	●	●	●	●	●	●	○	●
dataworld_10	○	●	●	●	●	○	○	●
dataworld_27	●	○	○	○	●	○	○	●
befchina_1	●	●	●	●	●	●	○	●
befchina_6	●	○	●	○	○	●	○	●
befchina_20	○	●	○	●	●	●	○	●
Bexis_24867	○	●	○	●	●	●	●	●
Bexis_25126	○	●	○	●	●	○	○	●
Bexis_25786	○	○	○	●	○	○	●	●
Bexis_27228	○	○	○	●	○	○	●	●

e.g., internal database “ID” columns or numerical columns without any explanation or context. We consider those datasets are impossible to annotate automatically and of little benefit to the community. Consequently, we decided to include only datasets containing essential categorical information. We selected 6 out of 32 dataset from data.world, 4 out of 15 from BExIS, and 3 out of 25 from BEFChina. data.world provides the most suitable datasets for STA, thus, it contributes about half of the datasets in BiodivTab. Our analysis of the collected data shows that, in addition to common challenges, real-world datasets feature unique characteristics. We enumerate the encountered challenges in our sample of datasets. We summarize their prevalence in Table 1.

- *Nested Entities*: more than one proper entity in a single cell, e.g., a chemical compound is combined with a unit of measurements.
- *Acronyms*: Abbreviations of different sorts are common, e.g., “Canna glauca”, a particular kind of flower, is often referred to as “C.glauca” or “Ca.glauca”.
- *Typos*: Data is predominantly collected manually by humans, so misspellings will occur, e.g., “Dead Leav” is used for “Dead Leaves”.
- *Numerical Data*: Most of the collected datasets describe the specimen by various measurements in numerical form.
- *Missing Values*: Data collected can be sparse and may include gaps, e.g., a column “super kingdoms” may consist of “unknown” values for the most part.
- *Lack of Context*: The collected data may barely provide any informative context for semantic annotations. e.g., a column with a missing or severely misspelled header.
- *Synecdoche*: Scientists may use a general entity as a short form to a more particular one, e.g., “Kentucky” is used instead of “Kentucky River”.
- *Specimen Data*: The collected datasets contain observations of particular specimens or groups, but do not pertain to the species as a whole.

3.2. Manual Annotation & Biodiversity Expert Revision

The annotation phase is the most time-consuming part of the benchmark creation since it included multiple rounds of revision. To ensure the quality of mappings, we manually annotated the selected tables with entities assembled from the live edition of Wikidata during September 2021, resulting in ground truth data for both CEA and CTA tasks. Concerning CEA, we have

Table 2

Which type would be correct for the given taxons?

Taxon	Type (A)	Type (B)
Bacteria (wd:Q10876)	superkingdom (wd:Q19858692)	taxon (wd:Q16521)
Actinobacteria (wd:Q130914)	phylum (wd:Q38348)	taxon (wd:Q16521)
Actinobacteria (wd:Q26262282)	class (wd:Q37517)	taxon (wd:Q16521)
Pseudonocardiales (wd:Q26265279)	order (wd:Q36602)	taxon (wd:Q16521)
Pseudonocardiaceae (wd:Q7255180)	family (wd:Q35409)	taxon (wd:Q16521)
Goodfellowiella (wd:Q26219639)	genus (wd:Q34740)	taxon (wd:Q16521)
Goodfellowiella coeruleoviolacea (wd:Q25859622)	species (wd:Q7432)	taxon (wd:Q16521)

marked possible candidate columns, typically those with categorical values, to annotate their cells. For each cell value, we assembled possible matches via Wikidata’s built-in search. We manually selected the most suitable matches to disambiguate the cells semantically if we found multiple matches. If we could not have chosen only one annotation, we pick all possible ones and consider them true matches. Thus, the provided ground truth contains all proper candidates for a given cell value. Biodiversity experts reviewed around 1/3 of the annotations. This revealed an error rate of about 1%. Because of the low error rate, the effort of this step outweighs the benefits. Thus, we have decided to continue annotating the remainder without further revisions.

We followed the same procedure for CTA. For categorical columns, we looked for a common type among column cells, taking into consideration the header value, to decide on the semantic type from Wikidata. Most of these columns are identified by the value of (wdt : P31, instance of) as the perfect annotation. However, finding such perfect annotation for taxon-related columns is not that easy. Since all taxon-related fields are instance of `taxon`. We believed it might not be distinguishable enough. In the biodiversity domain, experts are keen on more fine-grained modeling. E.g., species, genus, and class would be different types in their opinion. We established a simple one-question questionnaire for our biodiversity experts to select the perfect semantic type for a given taxonomic term as shown in Table 2. The first column shows the cell values with the corresponding mapping entities. The question is to select either which type is the most accurate, A, or B. We derive Type A from (wdt : P105, taxon rank) and Type B from (wdt : P31, instance of) in Wikidata. Based on their answers, the most fine-grained classification is (Type A); however, they consider (Type B) as a correct type as well. Thus, we have selected the perfect types for taxons through (wdt : P105, taxon rank). For numerical columns, most of them are identified by the column headers.

We maintain separate ground truth files to ease manual inspection, revision, and quality assurance for each table. So, “befchina_1”, e.g., is annotated by two such files: “befchina_1_CEA” and “befchina_1_CTA”. The structure of the ground truth files follows the format of SemTab challenge. 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*.

3.3. Data Augmentation

We further used data augmentation to increase the number of tables in our benchmark and reduce the human effort needed. In our context, we introduced challenges to the existing datasets based on our findings during the data collection and analysis phase, thus we rely on real-world challenges that we added programatically to increase the amount of the data. Table 3

Table 3

Data augmentation technique per dataset.

Dataset	Merge Cols	Separate Cols	Add Typos	Fix Typos	Disambiguate	Abbreviate	Increase Gap	Alter Cols	No. Files
dataworld_1	x3	-	-	-	-	-	x3	x1	7
dataworld_2	x3	-	x1	-	-	-	x1	-	5
dataworld_4	x4	-	-	x1	x1	x2	x1	-	9
dataworld_6	-	x1	-	-	-	-	-	-	1
dataworld_10	-	-	-	-	-	-	x1	-	1
dataworld_27	x1	-	x 2	-	-	-	-	-	3
befchina_1	x2	-	-	-	-	-	x1	-	3
befchina_20	x4	-	-	-	-	x1	-	-	5
Bexis_24867	-	-	-	-	-	x1	x2	-	3
Total									37

shows our used data augmentation techniques per dataset and the number of variations derived from it. In the following, we list techniques used and how they relate to the collected data issues:

- *Merge and Separate Columns* we either by introduced new nested entities or splited them up into separate columns.
- *Add and Fix Typos* we added noise to categorical cell values and, on rare occasions, fixed them.
- *Disambiguate* we replaced concepts with more accurate ones, e.g., the state is replaced by the river it stands for.
- *Abbreviate* we introduced more abbreviations especially with taxon-related values.
- *Alter Columns* we removed one or more data columns. This results in less informative and sparse datasets.

We managed to create the most variations from data.world since its datasets contain more categorical data that can be mapped to KG entities. Our data augmentation strategy increased the number of tables to 50 with less manual effort of the annotation.

3.4. CTA Ancestors Construction

To enable approximation of CTA F1, Precision and Recall scores [4], we provide an ancestors ground truth to our perfectly annotated types. The corresponding file is structured in a key-value format with keys representing the perfect annotation and values listing parent classes. We refer to those parents as *okay* classes.

Initially, we collected all unique column types from manually assigned perfect annotations. These are used to initialize a dictionary. Afterwards, we ran a sequence of three SPARQL queries sent to the public endpoint to retrieve related classes for each of them. For the first level, we query for direct types via (wdt : P31, instance of). We call them “E1”. For the second level, we query for further parent classes via (wdt : P279, subclass of) of the previous E1, resulting in “E2”. For the third and last level, we repeat the last process using the entities in E2, yielding “E3”. If the initial column type is a class (e.g., wd : Q60026969, unit of concentration) we skip the first step and only use the latter two. The resultant hierarchy consists of one perfect annotation with up to three levels of classes that are considered okay annotations. For taxon-related columns, we marked the (wdt : P105, taxonRank) as perfect annotation to follow the biodiversity experts’

recommendation. However, we have included (wd:Q16521, taxon) and (wd:Q21871294, living organism) as okay classes.

3.5. Assembly and Release

For publication, we anonymized the file names of tables to use unique identifiers using Python’s uuid functionalities. Subsequently, we aggregated the individual solutions of CEA and CTA-tasks into one file per task resulting in *CEA_biodivtab_2021_gt.csv* and *CTA_biodivtab_2021_gt.csv* respectively. We generated the corresponding “target-files” by removing the ground truth columns from these solution files. We provided anonymized tables alongside the target files to evaluate a particular system. The ground truth files alongside the dictionary for related classes, CTA-ancestors, are subsequently used to evaluate the results. Such way this follows the general approach of SemTab hiding the ground truth of STA-tasks from participants during the challenge. BiodivTab is awarded the first prize of IBM Research⁹ at the third round of 2021’s SemTab challenge [17] for its new challenges in CEA and CTA tasks.

3.6. DBpedia Ground Truth

In 2022¹⁰, we included annotations from DBpedia that are based on the Wikidata annotations in two ways: First, we exploited the link between Wikidata entities and corresponding Wikipedia pages. As there is a one-to-one correspondence between Wikipedia pages and DBpedia entities, we generated a Wikidata-DBpedia-mapping for them. Second, we extracted owl:sameAs mappings between Wikidata and DBpedia to complete our mapping from DBpedia itself. Despite these direct mappings appeared promising to begin with, they contain serious data quality issues. As of April 2022, *L-glutamic acid* (wd:Q26995161) is mapped to 1772 entities within the DBpedia graph using owl:sameAs. Thus, the resulting mappings were again manually verified to ensure the overall quality of the final DBpedia ground truth data. Generated types for CTA contained only instances/resources from DBpedia. During the manual verification, we further added classes from the DBpedia ontology as well. We attempted to replicate our approach from Wikidata using rdf:type and rdfs:subClassOf to retrieve the CTA-ancestors. However, some relations in the DBpedia ontology seemed unreasonable to us. For example, DBpedia at the time of writing contains a triple dbr:Species rdf:type dbo:MilitaryUnit. For these and other similar scenarios, we decided to not include an ancestor file for DBpedia.

4. Evaluation

In this section, we give a detailed overview of BiodivTab in terms of the size and content compared to existing benchmarks. In addition, we show the most and least frequent types of CTA. Finally, we demonstrate the application of our benchmark using the results of STA-systems during SemTab’s 2021 edition.

4.1. BiodivTab Characteristics

Table 4 summarizes the selected datasets in terms of their original and selected size, and the number of CEA and CTA mappings. For large datasets, e.g., *dataworld_4* and *dataworld_27*, we selected a subset of rows that retain the table characteristics. Most of the redundant species were dropped. Nevertheless, we kept the entire extent of BExIS datasets, including the redundant

⁹<https://www.research.ibm.com/>

¹⁰The new ground truth data from DBpedia is going to be used in SemTab 2022, thus we release a new benchmark after the conclusion of the challenge.

Table 4

Original and selected tables sizes, and entity and type mappings.

Dataset	Original Size		Selected Size		Mappings	
	Rows	Cols	Rows	Cols	CTA	CEA
dataworld_1	332	18	100	18	4	210
dataworld_2	37	25	37	8	8	226
dataworld_4	42 337	67	100	40	26	476
dataworld_6	271	6	100	6	4	103
dataworld_10	497	15	100	13	11	902
dataworld_27	95 368	12	100	12	5	398
befchina_1	7 553	16	145	16	3	294
befchina_6	26	4	26	4	2	53
befchina_20	787	45	99	43	28	304
Bexis_24867	151	13	151	13	9	159
Bexis_25126	4 906	35	4 906	14	6	9 816
Bexis_25786	2 001	39	2 001	21	5	4 017
Bexis_27228	1 549	8	1 549	8	3	4 646
Total					114	21 604
Avg.					8.8	1 661,8

Table 5

Most and least frequent semantic types in BiodivTab.

Most Frequent			Least Frequent		
Wikidata Id	Label	Freq.	Wikidata Id	Label	Freq.
wd:Q7432	Species	39	wd:Q8066	amino acid	1
wd:Q706	calcium	26	wd:Q11173	chemical compound	1
wd:Q577	year ¹¹	19	wd:Q60026969	unit of concentration	1
wd:Q677	iron	16	wd:Q2463705	Special Protection Area	1
wd:Q731	manganese	16	wd:Q1061524	intensity	1

entries, to achieve a good balance between the large tables and those with the reasonable length for STA-systems. The column mappings show the characteristic of specimen data, those columns with only local measurements and with local database names that could not be matched to the KG. For example, only 4 out of 18 columns in *dataworld_1* could be matched to KG-entities.

Figure 3 shows the domains distribution of the 83 unique semantic types in the CTA-solutions. Approximately two-thirds of these types belong to the biodiversity domain. The distinction into the biodiversity-related, general domain, and mixed types was made according to the definitions introduced in [21, 22]. General domain types include, e.g., visibility, scale, cost, and airport. Mixed domain types contain examples like river, temperature, or sex of humans. Biodiversity-related types include taxon, chemical compounds, and soil type. In addition, Table 5 provides a list of most and least frequent semantic types in BiodivTab. Species (wd:Q7432) is the most frequent type, which reflects its importance in biodiversity research.

**Figure 3:** Domain distribution in BiodivTab benchmark.

Table 6 shows both data sources and target KGs or resource for BiodivTab and existing benchmarks. The three editions of SemTab from 2019 to 2021 [6, 7, 8] used both Wikidata and Wikipedia[23] as table sources. However, the target KGs varies between using DBpedia,

¹¹Calendar year, wd:Q3186692, is equivalent to year, wd:Q577.

Table 6

Data sources for existing benchmarks and their corresponding targets. Entries for SemTab are aggregated over all rounds each.

Dataset	Data Source	Target Annotation
SemTab 2019	Wikidata, Wikipedia	DBpedia
SemTab 2020	Wikidata, Wikipedia	Wikidata
SemTab 2021	Wikidata, Wikipedia	Wikidata, DBpedia
T2Dv2	WebTables	DBpedia
Limaye	Wikipedia	DBpedia
GitTables	GitHub	DBpedia, Schema.org
BiodivTab	BExIS, BEFChina, data.world	Wikidata, DBpedia

Wikidata, or both. T2Dv2 [11] and Limaye [10] use the WebTables [24] and Wikipedia as their data sources respectively while having annotations from DBpedia. GitTables [25] and the adapted version [9] for SemTab 2021 challenge, leverages GitHub as a table source and provide annotations from DBpedia and schema.org. Unlike all the previous benchmarks, BiodivTab uses domain-specific data portals, as table sources. It provides Wikidata annotations like SemTab 2020 and 2021.

Table 7 shows a comparison between BiodivTab and existing benchmarks in terms of the average number of rows, columns, and cells. It also gives an overview of the targets for CEA, CTA, and CPA. BiodivTab is the smallest in terms of the number of tables. However, BiodivTab has the maximum average number of columns, and average number of rows except for SemTab 2021, Round 1, and BioTables in Round 2. This poses an additional challenge for STA systems. For CTA targets, BiodivTab is a middle point among the existing benchmarks.

4.2. Applications

Table 8 shows the scores from SemTab2021 top participants on BiodivTab and HardTables during Round 3. Scores have been published by the organizers of SemTab2021 [17]. The details about the mentioned systems using BiodivTab are beyond the scope of this paper. For BiodivTab, CEA has maximum F1-score by JenTab [20] of 60.2%, while the CTA has a maximum score with 59.3% by KEPLER [26]. In contrast, for the synthetic dataset, HardTables, DAGOBAB achieved the maximum F1-score 97.4%, and 99% for CEA, and CTA respectively. These results show that annotating real-world, domain-specific tables is far from solved by state-of-the-art STA-systems. This underlines the importance of benchmarks like BiodivTab further to foster the transfer of academic projects to real-world applications.

4.3. Availability and Long-Term Plan

Resources should be easily accessible to allow replication and reuse. We follow the FAIR (Findable, Accessible, Interoperable, and Reusable) guidelines to publish our contributions [14]. We release our dataset [29] in such a way that researchers in the community can benefit from it. In addition, we release the code [30] that was used to augment the data, assemble, and reconcile the benchmark. Our dataset and code are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) License and Apache License 2.0 respectively.

5. Conclusions and Future Work

We introduced BiodivTab, the first biodiversity tabular benchmark for Semantic Table Annotation tasks. It consists of a collection of 50 tables. BiodivTab as created manually by annotating 13

Table 7

Comparison with existing benchmarks. *ST19 - ST21* (SemTab editions). *_*W* and *_*D* use Wikidata and DBpedia as targets. *ST21-H2*, and *H3* are HardTables for Round 2 and 3 during SemTab2021. *ST21-Bio* is BioTables at SemTab2021 Round 2. *ST21-Git* is the published version of GitTables during SemTab2021 Round 3.

Dataset	Tables	Avg. Rows (\pm Std Dev.)	Avg. Cols (\pm Std Dev.)	Avg. Cells (\pm Std Dev.)	CEA	CTA	CPA
ST19-R1	64	142 \pm 139	5 \pm 2	696 \pm 715	8,418	120	116
ST19-R2	11,924	25 \pm 52	5 \pm 3	124 \pm 281	463,796	14,780	6,762
ST19-R3	2,161	71 \pm 58	5 \pm 1	313 \pm 262	406,827	5,752	7,575
ST19-R4	817	63 \pm 52	4 \pm 1	268 \pm 223	107,352	1,732	2,747
ST20-R1	34,294	7 \pm 4	5 \pm 1	36 \pm 20	985,110	34,294	135,774
ST20-R2	12,173	7 \pm 7	5 \pm 1	36 \pm 18	283,446	26,726	43,753
ST20-R3	62,614	7 \pm 5	4 \pm 1	23 \pm 18	768,324	97,585	166,633
ST20-R4	22,390	109 \pm 11,120	4 \pm 1	342 \pm 33,362	1,662,164	32,461	56,475
ST21-R1_W	180	1,080 \pm 2,798	5 \pm 2	4125 \pm 10947	663,655	539	NA
ST21-R1_D	180	1,080 \pm 2,798	4 \pm 2	3,952 \pm 10,129	636,185	535	NA
ST21-H2	1,750	17 \pm 8	3 \pm 1	55 \pm 32	47,439	2,190	3,835
ST21-Bio	110	2,448 \pm 193	6 \pm 1	14,605 \pm 2,338	1,391,324	656	546
ST21-H3	7,207	8 \pm 5	2 \pm 1	20 \pm 15	58,948	7,206	10,694
ST21-Git	1,101	58 \pm 95	16 \pm 12	690 \pm 1,159	NA	2,516	NA
ST21-Git	1,101	58 \pm 95	16 \pm 12	690 \pm 1,159	NA	720	NA
T2Dv2	779	85 \pm 270	5 \pm 3	359 \pm 882	NA	237	NA
Limaye	428	24 \pm 22	2 \pm 1	51 \pm 50	NA	84	NA
BiodivTab_W	50	259 \pm 743	24 \pm 13	4,589 \pm 10,862	33,405	614	NA
BiodivTab_D	50	259 \pm 743	24 \pm 13	4,589 \pm 10,862	33,405	569	NA

Table 8

SemTab2021 top participants' scores for BiodivTab and HardTables 3 benchmarks. F1 - F1 Score, Pr - Precision, R - Recall, AF1, APr, and AR - Approximate version of F1 Score, Precision, and Recall respectively. Highest scores are in **bold**.

System	BiodivTab						HardTables 3					
	CEA			CTA			CEA			CTA		
	F1	Pr	R	AF1	APr	AR	F1	Pr	R	AF1	APr	AR
MTab [19]	0.522	0.527	0.517	0.123	0.282	0.079	0.968	0.968	0.968	0.984	0.984	0.984
Magic [27]	0.142	0.192	0.112	0.1	0.253	0.063	0.641	0.721	0.577	0.687	0.687	0.688
DAGOBAAH [18]	0.496	0.497	0.495	0.381	0.382	0.38	0.974	0.974	0.974	0.99	0.99	0.99
mantisTable [28]	0.264	0.785	0.159	0.061	0.076	0.051	0.959	0.984	0.935	0.965	0.973	0.958
JenTab [20]	0.602	0.611	0.539	0.107	0.107	0.107	0.94	0.94	0.939	0.942	0.942	0.942
KEPLER [26]	NA	NA	NA	0.593	0.595	0.591	NA	NA	NA	0.244	0.279	0.217

tables from real-world biodiversity datasets and adding 37 more tables by augmenting them with noise based on challenges that are commonly observed in the domain. The target knowledge graphs for annotations are Wikidata and DBpedia. An evaluation during SemTab2021 showed that current state-of-the-art systems still struggle with the challenges posed. This highlights BiodivTab's importance for further development in the field. BiodivTab itself and the code used to create it are publicly available.

Future Work We see multiple directions to continue this work. We plan to include more biodiversity tables from other projects to cover a broader domain spectrum. We also plan to apply further quality checks of the annotations like multiple-annotators annotation and validation via the interrater agreement. In addition, we plan to provide ground truth data

from other knowledge graphs, particularly domain-specific ones. Moreover, we analyze the performance of STA-systems on the BiodivTab.

Acknowledgments

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 thank our biodiversity experts Cornelia Fürstenau and Andreas Ostrowski for feedback and validation of the created annotations.

References

- [1] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* 57 (2014) 78–85. doi:10.1145/2629489.
- [2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: *The semantic web, 2007*, pp. 722–735.
- [3] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, SemTab 2019: Resources to benchmark tabular data to knowledge graph matching systems, in: *European Semantic Web Conference, Springer, 2020*, pp. 514–530.
- [4] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona, Results of SemTab 2020, in: *CEUR, volume 2775, 2020*, pp. 1–8.
- [5] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, V. Cutrona, Results of SemTab 2021, in: *CEUR Workshop Proceedings, 2021*.
- [6] O. Hassanzadeh, V. Efthymiou, J. Chen, E. Jiménez-Ruiz, K. Srinivas, SemTab 2019: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching - 2019 Data Sets, 2019. doi:10.5281/zenodo.3518539.
- [7] O. Hassanzadeh, V. Efthymiou, J. Chen, E. Jiménez-Ruiz, K. Srinivas, SemTab 2020: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Data Sets, 2020. doi:10.5281/zenodo.4282879.
- [8] O. Hassanzadeh, V. Efthymiou, J. Chen, E. Jiménez-Ruiz, K. Srinivas, SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Data Sets, 2021.
- [9] M. Hulsebos, Ç. Demiralp, P. Groth, GitTables: A Large-Scale Corpus of Relational Tables, *arXiv preprint arXiv:2106.07258* (2021).
- [10] G. Limaye, S. Sarawagi, S. Chakrabarti, Annotating and searching web tables using entities, types and relationships, *Proceedings of the VLDB Endowment* 3 (2010) 1338–1347.
- [11] O. Lehmberg, D. Ritze, R. Meusel, C. Bizer, A large public corpus of web tables containing time and context metadata, in: *Proceedings of the 25th International Conference Companion on World Wide Web, 2016*, pp. 75–76.
- [12] V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, M. Palmonari, Tough Tables: Carefully Evaluating Entity Linking for Tabular Data, 2020. doi:10.5281/zenodo.4246370.
- [13] U. Küster, B. König-Ries, Towards standard test collections for the empirical evaluation of semantic web service approaches, *International Journal of Semantic Computing* 2 (2008) 381–402.
- [14] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR Guid-

- ing Principles for scientific data management and stewardship, *Scientific data* 3 (2016) 1–9.
- [15] L. M. Gadelha Jr, P. C. de Siracusa, E. C. Dalcin, L. A. E. da Silva, D. A. Augusto, E. Krempser, H. M. Affe, R. L. Costa, M. L. Mondelli, P. M. Meirelles, et al., A survey of biodiversity informatics: Concepts, practices, and challenges, *Wiley Interdisciplinary Reviews* 11 (2021) e1394.
 - [16] 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 co-located with the 20th International Semantic Web Conference (ISWC 2021)*, Virtual conference, October 27, 2021, volume 3103 of *CEUR*, CEUR-WS.org, 2021, pp. 13–18.
 - [17] V. Cutrona, J. Chen, V. Efthymiou, O. Hassanzadeh, E. Jiménez-Ruiz, J. Sequeda, K. Srinivas, N. Abdelmageed, M. Hulsebos, D. Oliveira, et al., Results of SemTab 2021, *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching* 3103 (2022) 1–12.
 - [18] V.-P. Huynh, J. Liu, Y. Chabot, T. Labbé, P. Monnin, R. Troncy, DAGOBAB: Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data., in: *SemTab@ ISWC*, 2021.
 - [19] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, SemTab 2021: Tabular Data Annotation with MTab Tool., in: *SemTab@ ISWC*, 2021.
 - [20] N. Abdelmageed, S. Schindler, JenTab Meets SemTab 2021’s New Challenges., in: *SemTab@ ISWC*, 2021.
 - [21] N. Abdelmageed, A. Algergawy, S. Samuel, B. König-Ries, BiodivOnto: Towards a core ontology for biodiversity, in: *European Semantic Web Conference (ESWC)*, Springer, 2021, pp. 3–8.
 - [22] F. Löffler, V. Wesp, B. König-Ries, F. Klan, Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?, *PloS one* 16 (2021) e0246099.
 - [23] C. S. Bhagavatula, T. Noraset, D. Downey, Methods for exploring and mining tables on wikipedia, in: *Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics*, 2013, pp. 18–26.
 - [24] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, Y. Zhang, Webtables: exploring the power of tables on the web, *VLDB Endowment* 1 (2008) 538–549.
 - [25] M. Hulsebos, Çağatay Demiralp, P. Demiralp, Gittables for SemTab 2021 - cta task, 2021. doi:10.5281/zenodo.5706316.
 - [26] W. Baazouzi, M. Kachroudi, S. Faiz, Kepler-aSI at SemTab 2021., in: *SemTab@ ISWC*, 2021.
 - [27] B. Steenwinckel, F. De Turck, F. Ongenae, MAGIC: Mining an Augmented Graph using INK, starting from a CSV., in: *SemTab@ ISWC*, 2021.
 - [28] R. Avogadro, M. Cremaschi, MantisTable V: A novel and efficient approach to Semantic Table Interpretation., in: *SemTab@ ISWC*, 2021.
 - [29] N. Abdelmageed, S. Schindler, B. König-Ries, fusion-jena/BiodivTab, 2022. doi:10.5281/zenodo.6461556.
 - [30] N. Abdelmageed, S. Schindler, B. König-Ries, fusion-jena/biodivtab: Benchmark data and code, 2021. doi:10.5281/zenodo.5749340.