MammoTab 25: A Large-Scale Dataset for Semantic Table Interpretation - Training, Testing, and Detecting Weaknesses

Marco Cremaschi[1](✉), Federico Belotti[1], Jennifer D'Souza[2], and Matteo Palmonari[1]

[1] Department of Informatics, Systems and Communication - DISCo, University of Milano-Bicocca, Milan, Italy
{marco.cremaschi,federico.belotti,matteo.palmonari}@unimib.it

[2] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
jennifer.dsouza@tib.eu

Abstract. The paper presents *MammoTab 25*, a new dataset comprising approximately 838930 Wikipedia tables extracted from over 63 million English Wikipedia pages and semantically annotated through Wikidata. Each table in *MammoTab 25* is accompanied by fine-grained metadata, including column typing, NIL flags, and statistics, and by four prompt templates, making the resource simultaneously suitable for training, fine-tuning, and stress-testing Large Language Models (LLMs). *MammoTab 25* covers, in a single benchmark, all key challenges for the semantic interpretation of tables, such as disambiguation issues, homonymy and acronym presence, NIL-mentions, and large web-table sizes; the tags attached to every table let researchers isolate and diagnose specific failure cases with precision. The corpus is delivered with an open-source pipeline that can be rerun on future Wikipedia dumps, ensuring long-term sustainability and up-to-date annotations. *MammoTab 25* already supports, and will continue to support, a public leaderboard that evaluates the Semantic Table Interpretation (STI) capabilities of state-of-the-art and upcoming LLMs, providing the community with a live yardstick of progress.

Resource Type: Dataset
License: GNU Affero General Public License v3.0
DOI: https://doi.org/10.5281/zenodo.16562700
URL: https://github.com/unimib-datAI/mammotab/
Website/Documentation:
https://unimib-datai.github.io/mammotab-docs/

Keywords: Semantic Table Interpretation · Large Language Models · Benchmark Dataset · Benchmarking Initiative

D. Garijo et al. (Eds.): ISWC 2025, LNCS 16141, pp. 131–148, 2026.
https://doi.org/10.1007/978-3-032-09530-5_8

1 Introduction

The explosion of digital information has dramatically increased the volume of structured and unstructured data, with tabular formats remaining a core means of organising knowledge, from ancient bookkeeping to modern data-centric workflows. One of the earliest examples dates to around 2500 BC, when Egyptian naval inspector Merer recorded his activities on a papyrus table [27]. While tables have long supported commerce and science, their widespread dissemination accelerated with the advent of the Open Data movement. Today, vast, high-quality collections spanning finance, mobility, tourism, sports, and cultural heritage are publicly available [21]. Web crawlers extract billions of HTML tables annually, millions of which pass stringent quality filters[1]. Wikipedia alone hosts millions of tables across diverse domains, underscoring the central role of tabular data in knowledge dissemination. This proliferation of tabular data necessitates advanced methods for its interpretation, leading to the crucial task of STI.

STI, the task of enriching tables by linking elements (cells, columns, and rows) to semantic tags (entities, classes, and properties) from Knowledge Graphs (KGs) such as Wikidata and DBpedia, involves three subtasks: cell-to-entity annotation (CEA), column-to-class annotation (CTA), and column-pair-to-property annotation (CPA) [7]. STI, also called table interpretation or table understanding, has gained traction across Semantic Web, Data Management, AI, and NLP communities [7,10,12,13], and is central to KG enrichment [8,15,22,28]. Since 2019, the SemTab challenge [7,12,13][2] has served as a benchmark venue for STI research. Notably, CEA—akin to entity linking or reconciliation has applications beyond STI, including a W3C-endorsed reconciliation API for cross-service interoperability[3].

In parallel, STI has become a benchmark for evaluating the reasoning abilities of Large Language Models (LLMs). LLMs are pre-trained on vast corpora combining unstructured text, semi-structured content, and HTML tables, enabling them to internalise lexical context and latent schema regularities necessary for aligning table cells with entities, types, and relations. Their large context windows support joint reasoning over table structure, surrounding text, and background knowledge, capabilities that conventional neural architectures, often optimised for narrow tasks, struggle to replicate. Moreover, LLMs can be prompted or fine-tuned with only a handful of annotated examples, enabling rapid adaptation to new domains without the need for extensive retraining. These strengths have led to strong performance on STI benchmarks across various tasks, including entity linking, type annotation, and relation extraction, frequently surpassing task-specific systems. As a result, LLMs are increasingly viewed as robust, general-purpose components in data processing pipelines, underscoring the need

[1] Common Crawl curates a freely accessible repository of web crawl data: commoncrawl.org.

[2] SemTab Website - www.cs.ox.ac.uk/isg/challenges/sem-tab/.

[3] Reconciliation Service API v0.2 - www.w3.org/community/reports/reconciliation/FINAL-specs-0.2-20230410/.

for STI benchmarks that accurately reflect the real-world complexity, diversity, and ambiguity of web tables.

Despite the development of several STI datasets (Gold Standards), there is a general acknowledgement that they often cover only a subset of the characteristics of web tables (*e.g.*, small tables with few mentions, which are easy to annotate semantically [30]). In response to this gap, we developed *MammoTab 25*[4], a dataset comprising 838930 tables extracted from 63 million English Wikipedia pages and annotated through Wikidata. We present this as a novel resource in the community. Crucially, *MammoTab 25* is curated as a Gold Standard resource for CEA reconciliation and serves as a Silver Standard benchmark for CTA and CPA, a distinction examined in detail later in this paper. The tables in *MammoTab 25* systematically cover the key challenges of STI (cf. Section 3), enabling precise diagnosis of where and why different approaches may fail. Each table is complemented by four prompt templates designed to support fine-tuning and evaluating LLM-based models. Thanks to its large scale and rich structure, *MammoTab 25* is a valuable benchmark for developing and assessing STI methods, especially those that require extensive training data, such as LLM-based Entity Linking (EL) systems.

Compared to its predecessor [19], *MammoTab 25* contains fewer tables (838930 *vs* 980254) but features higher-quality annotations, thanks to an updated version of *LamAPI*[5] (*La*bel *m*atching *API*) [2,3], a comprehensive IR-based Entity Retrieval (ER) tool enhanced with type-based filtering. Additional data-cleaning procedures were applied, resulting in the removal of noisy tables[6]. Rich metadata were also added per table, detailing STI-relevant challenges. A preliminary release (v2-alpha) powered Round 2 of the "LLM vs STI" track at SemTab 2024[7] where three systems used it, while a curated subset of 870 tables with 84907 distinct mentions has been used as the primary dataset for SemTab 2025[8]. The dataset's continued use in these challenges underscores its robustness and practical value.

The remainder of the paper is structured as follows: Sect. 2 defines the STI task and its challenges. Section 3 reviews related datasets. Section 4 presents *MammoTab 25*, with its characteristics in Sect. 4.1 and implementation in Sect. 4.2. Section 5 reports evaluation results using state-of-the-art STI methods. Section 6 concludes and outlines future work.

[4] The name "MammoTab" combines mammoth, evoking something exceptionally large, with tab (short for table). The intention is to emphasise that the collection is a "mammoth of tables": a very large-scale, richly annotated corpus.

[5] LamAPI API docs – lamapi.datai.disco.unimib.it.

[6] List of data-cleaning rules – unimib-datai.github.io/mammotab-docs/docs/data-cleaning.

[7] SemTab 2024 "LLM vs STI" track – sem-tab-challenge.github.io/2024/tracks/sti-vs-llm-track.html.

[8] SemTab 2025 – sem-tab-challenge.github.io/2025/.

2 Semantic Table Interpretation: Tasks Definition

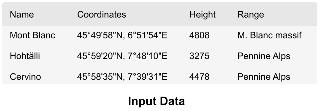
In the state-of-the-art (SOTA), it is possible to identify different conceptualisations of the table annotation problem. Considering the formalisation proposed in the SemTab challenge as one of the formulations that has obtained the most significant consensus, we can define the STI as follows.

Given a *relational table T* (Fig. 1), a *Knowledge Graph* (entities + statements), and its *Ontology* (types + properties) (Fig. 2), *T* is considered annotated when:

- each column is associated with one or more KG-types [Column-Type Annotation (CTA)] (*e.g.*, the column *Name* is annotated with the type `Mountain`; the column *Height* is annotated with datatype `xsd:integer`);
- each cell in "entity columns" is annotated with a KG-entity or with *NIL* (if not in the KG) [Cell-Entity Annotation (CEA)] (*e.g.*, the cell *Le Mont Blanc* is annotated with `Mont_Blanc`; the cell *Hohtälli* is annotated with *NIL* since it has not yet been included in the KG);
- some pair of columns is annotated with a binary KG-property [Columns-Property Annotation (CPA)] (*e.g.*, the pairs composed by the columns *Name* and *Height* are annotated with `dbo:elevation`).

The result of the annotation process can be seen in Fig. 3. In the example, the annotations can also generate new elements within the Knowledge Graph (KG) (*i.e.*, the entity `Hohtälli`). This conceptualisation covers most of the proposed definitions by generalising some aspects (*e.g.*, consideration of NILs and selection of column-pairs to annotate).

As inferred from the example, the key step of STI is *finding links* between strings (or mentions) present in the table and the entities they reference in a background KG. This task is referred to as Entity Linking (EL), borrowing the term from Language Processing (NLP) where EL algorithms have been studied extensively and are used today in countless real-world applications. EL enables the integration, enrichment, and extension of the data and/or KGs.



Name	Coordinates	Height	Range
Mont Blanc	45°49'58"N, 6°51'54"E	4808	M. Blanc massif
Hohtälli	45°59'20"N, 7°48'10"E	3275	Pennine Alps
Cervino	45°58'35"N, 7°39'31"E	4478	Pennine Alps

Input Data

Fig. 1. Example of a well-formed relational table.

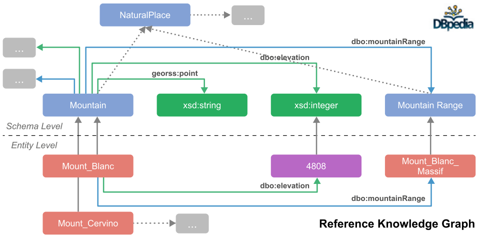


Fig. 2. A sample of Knowledge Graph.

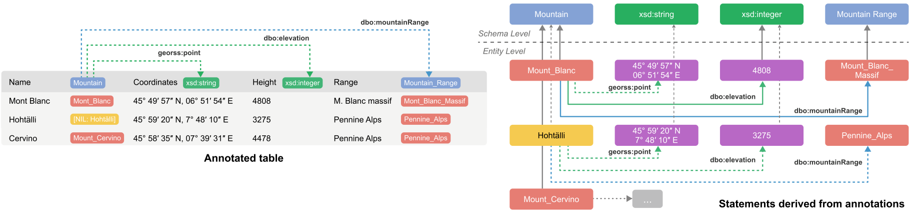


Fig. 3. Example of an annotated table.

3 Comparison of *MammoTab 25* and Other Datasets

Several datasets have been developed to support research on STI, including T2Dv2 [24], WebTableStitching [24], Limaye [18], LimayeAll [32], Limaye200 [32], MusicBrainz [32], IMDB [32], Taheriyan [26], Tough Table [6], SOTAB [16], Wikary [20], GitTables [11], REDTab [25], TURL [9], BiodivTab [1], TSOTSACorpus [14], and SemTab. While these datasets provide valuable benchmarks, they often exhibit essential limitations, especially when training and evaluating approaches based on LLMs. In most cases, they cover only a narrow subset of the characteristics of web tables, such as small size, few mentions, or semantically unambiguous content. This restricts their utility in assessing the generalisation and robustness of modern LLM-based systems. Recently, *TableInstruct* [31] was introduced as a comprehensive instruction-tuning dataset for tables, encompassing diverse real-world tables and tasks such as CEA, CTA, and CPA. However, the "Table Interpretation" split[9] is derived from TURL [9] and thus inherits its limitations (see Table 1). Furthermore, *TableInstruct* prioritises task diversity for LLM training without directly addressing any specific challenges, as is the case with *MammoTab 25*. Table 1 summarises key statistics for the gold standards[10].

Creating a successful STI approach requires cautious planning and skilful management of various table components. Furthermore, the annotation process poses several challenges that must be carefully addressed to ensure accuracy and efficacy:

1. *Distinguishing between column types*: In a table, there can be columns that contain references to named-entities (NE-column) and columns that include strings, numbers, dates, and, in general, values that are instances of specific data types, which we refer to as literals (L-column); distinguishing between the two types of columns is crucial to support the annotation process.
2. *Heterogeneity of domains*: A table may cover information that refers to very different domains (*e.g.*, Geography *vs* Sports).

[9] "Table Interpretation split" denotes the "Task Category" labels used in Table 1 of [31].

[10] An interactive version of the table is available at: unimib-datai.github.io/sti-website/datasets/.

Table 1. Statistics for the most common datasets. '—' indicates the value is unknown. The Challenges column lists, for each dataset, the benchmark challenges described in Sect. 3 that it covers; an index (1–10) appears when the dataset includes instances for that particular challenge.

GS		Tables	Cols (min | max | avg)	Rows (min | max | avg)	Classes	Entities	Pred.	KG	Challenges
T2Dv2 [24]		234	1,2K (1 | 30 | 4,52)	2,8K (1 | 5K | 84,55)	39	—	154	DBpedia	1 2 3 4 5 8
WebTableStitching [24]		50	300 (6 | 6 | 6)	717 (3 | 83 | 14,84)	9	400	6	DBpedia	2 10
Limaye [18]		6,5K	—	—	747	143K	90	Wikipedia Yago	1 2 3 4 5 8
LimayeAll [32]		6,3K	28,5K	136K	—	227K	—	Freebase	1 2 3 4 5 8
Limaye200 [32]		200	903	4,1K	615	—	361	Freebase	1 2 3 4 5 8
MusicBrainz [32]		1,4K	9,8K	—	—	93,3K	7K	Freebase	1 5 8
IMDB [32]		7,4K	7,4K	—	—	92,3K	—	Freebase	1 5 8
Taheriyan [26]		29	2,5K (3 | 71,3K | 529K)	16K (1 | 13,8K | 957)	—	—	—	Schema.org	2 4 5 8
Tough Table (2T) [6]		180	802 (1 | 8 | 4,46)	194K (6 | 15,5K | 108K)	540	667K	0	Wikidata DBpedia	1 2 3 4 5 7 8
SOTAB [16]		108K	—	—	91	—	176	Schema.org	1 2 3 4 5 8
Wikary [20]		81,7K	22,5K	63,9K	—	30,6K	188	Wikidata	1 2 3 4 5 10
GitTables [11]		962K	11,5M	13,6M	2,4K	—	—	Schema.org DBpedia	1 2 3 5 8 10
REDTab [25]		9K	44,6K (1 | 11 | 4,86)	148K (1 | 353 | 17,09)	70	—	23	Music Literature	1 4 5 8
TURL [9]		484K	2,8M	7,9M	—	1,2M	—	Wikidata DBpedia	2 3 5 8 9 10
BiodivTab [1]		50	1,2K (1 | 43 | 23,96)	12,9K (26 | 4,9K | 261)	84	1,2K	—	Wikidata	1 4 5 7 8
TSOTSACorpus [14]		16K	—	—	200	60K	—	Food Data	1 4 5
SemTab2019	**R1**	64	320 (3 | 14 | 5,05)	9K (7 | 586 | 143)	120	8,4K	116	DBpedia	1 2 3 4 5 8
	R2	11,9K	59,6K (1 | 51 | 5,55)	29,8K (1 | 1,5K | 27,06)	14,8K	464K	6,7K		
	R3	2,1K	10,8K (4 | 8 | 4,51)	153K (6 | 207 | 71,69)	5,7K	407K	7,6K		
	R4	817	3,3K (4 | 8 | 4,36)	51,4K (6 | 198 | 63,73)	1,7K	107K	2,7K		
SemTab2020	**R1**	34,3K	170K (4 | 8 | 4,96)	249K (5 | 16 | 8,27)	136K	985K	136K	Wikidata	1 2 3 4 5 8
	R2	12,1K	55,9K (4 | 8 | 4,6)	84,9K (5 | 16 | 7,97)	438K	283K	43,8K		
	R3	62,6K	229K (3 | 7 | 3,66)	397K (3 | 16 | 7,34)	167K	768K	167K		
	R4	22,4K	79,6K (1 | 8 | 3,55)	670K (6 | 15,5K | 30,94)	32,5K	1,7M	56,5K		
SemTab2021	**R1**	180	802 (1 | 8 | 4,46)	194K (6 | 15,5K | 1,08K)	539	667K	56,5K	Wikidata DBpedia	1 2 3 4 5 8
	R2	1,7K	5,6K (2 | 7 | 3,19)	29,3K (5 | 58 | 17,73)	2,1K	47,4K	3,8K		
	R3	7,2K	17,9K (2 | 5 | 2,48)	58,9K (5 | 21 | 9,18)	7,2K	58,9K	10,7K		
SemTab2022	**R1 HT Test**	3,6K	9,4K (2 | 5 | 2,56)	21K (4 | 8 | 5,68)	416	26,2K	360	Wikidata	1 2 3 4 5 8
	R1 HT Valid	200	519 (2 | 5 | 2,59)	1,1K (4 | 8 | 5,74)	142	1,4K	140		
	R2 HT Test	4,6K	11,9K (2 | 5 | 2,56)	26K (4 | 8 | 5,57)	405	22K	332		
	R2 HT Valid	457	1,17K (2 | 5 | 2,55)	2,5K (4 | 8 | 5,54)	164	1,98K	133		
	R2 2T	36	155K (1 | 8 | 4,3)	24K (7 | 8,2K | 683)	97	81,1K	—		
	R3 Biodiv	50	1,2K	12,9K	43	1,5K	—		
	R3 GitTables	7,6K	198K	841K	6,2K 4,4K 1K	—	—		
SemTab2023	**R1 Wikidata Test**	9,9K	24,9K (2 | 4 | 2,51)	56K (3 | 11 | 5,65)	1 809	64,5K	556	Wikidata	1 2 3 4 5 7 8
	R1 Wikidata Valid	500	1,2K (2 | 4 | 2,46)	3,47K (3 | 11 | 6,9)	194	4,2K	177		
MammoTab 22 (v.1) [19]		980K	5,6M (1 | 500 | 5,75)	23,3M (3 | 14,4k | 23,85)	3.3K	28M	—	Wikidata	1 2 3 4 5 7 8 10
MammoTab 25 (v.2)		838K	4,2M (1 | 676 | 5)	20,9M (3 | 9,9k | 25)	4.6K	26M	—	Wikidata	1 2 3 4 5 7 8 10

3. *Diversity of data*: The specificity of the table content may vary significantly (from a table with basic information about most famous mountains, like in Fig. 1, to a table that contains the composition of mountain rocks[11]).
4. *Limited contextual information*: If compared with similar interpretation and disambiguation tasks on the textual document, the presence of contextual

[11] List of rock formations - en.wikipedia.org/wiki/List_of_rock_formations.

clues to support the interpretation and annotation of table elements may be limited and very diverse depending on the data sources. In this challenge, we can identify different cases: i) tables with/without headers; ii) tables with one or few NE-columns; iii) tables made up mainly of L-columns; iv) tables with/without metadata or external context (*e.g.*, a surrounding text in a web page, a caption). Providing solutions that deal with this scarcity and diversity is quite challenging.

5. *Matching tabular values against the KG*: Matching the values in the table to the data in the KG is a typical approach for collecting evidence to interpret the table. The table mentions often differ from the entity's label in a KG. Related to the named-entities, these differences may be due to: i) use of acronyms; ii) use of aliases; iii) presence of typos. For example, *High Tauern* refers to the *Johannisberg mountain (High Tauern)* entity in DBpedia[12]. Regarding literals, the difference between the values in the tables and the values in the KG may be due to: i) approximate values; ii) different units of measurement; iii) outdated data. For instance, the height of mountains may be different, *e.g.*, because outdated, measured differently and so on.
6. *Disambiguation of named-entities*: The KG may contain many entities with similar or even equal names (homonyms) that may belong to different or the same type. For example, the mention *Mont_Blanc* in Fig. 1, which refers to the famous mountain located on the French-Italian border[13], matches labels of different entities associated with different types, including a tunnel, a poem, and a dessert[14]; the same mention matches also the label of a mountain in the Moon[15].
7. *Identification of Not In Lexicon (NIL)-mentions*: The approach must also consider the NIL-mentions, *i.e.*, strings that refer to entities for which a representation has not yet been created within the KG. The mention *Hohtälli* in Fig. 1, a mountain of the Swiss Pennine Alps, does not currently have a counterpart in Wikidata. For this reason, it must be annotated as NIL.
8. *Choosing the most appropriate types*: The KG may contain hundreds or thousands of types to choose from for annotating columns; entities are classified with multiple types, which may reflect different levels of specificity (*e.g.*, if we consider the subclass of the relation between types, Mont Blanc is a mountain, a summit, a pyramidal peak, an elevation, a landform, a geographical feature, a geographic location and more in Wikidata). The classification may be more or less complete depending on the specific entity; deciding which type or which set of types better describes a set of entities in a column is not trivial.
9. *Choosing the most appropriate properties*: The KG may contain hundreds or thousands of properties to choose from for annotating column-pairs; several column-pairs could be potentially related, and several properties exist in KG

[12] Johannisberg (High Tauern) - dbpedia.org/page/Johannisberg_(High_Tauern).

[13] Mont Blanc - dbpedia.org/page/Mont_Blanc.

[14] Mont Blanc (disambiguation) - en.wikipedia.org/wiki/Mont_Blanc_(disambiguation).

[15] Mont Blanc (Moon) - dbpedia.org/page/Mont_Blanc_(Moon).

that have similar meanings (also, in this case, properties can be organised into taxonomies and have different levels of specificity) [23]; choosing the column-pairs that should be annotated with a property and selecting the property that more precisely describes the semantics of a column-pair is not trivial either.

10. *Amount and shape of data*: Depending on the application scenarios, it may be necessary to process a large number of small tables or very large tables [4], which may imply different constraints on the approaches or introduce slightly different challenges; more scenarios may also become more relevant in the future, such as processing streaming data that can be formatted as tabular data.

The previous challenges cover various scenarios, and combining different challenges significantly increases the potential issues that an STI approach must consider. The analysis of the previously mentioned datasets reveals that they address only a limited subset of these challenges. Most datasets focus on the disambiguation of heterogeneous domains, while fewer consider issues related to data distribution. Only in recent datasets, such as those from SemTab 2022, there has been an effort to tackle challenges like selecting the most appropriate types (with multiple score for "correct annotation" and "ok annotation") and identifying NIL mentions, along with handling variations in data quantity, structure, and the presence of typos and acronyms.

L-columns (columns that contain literals) are generally overlooked, meaning that selecting the most appropriate properties is not considered. Additionally, aliases are often disregarded and then not annotated. Moreover, contextual information, such as headers and metadata, is absent in most datasets.

As explained later in the paper, the *MammoTab 25* generation script exposes configurable thresholds that let users tailor the dataset's properties, for instance, by injecting typos, acronyms, aliases, or approximating numeric values[16].

4 The *MammoTab 25* Dataset Resource

As anticipated, *MammoTab 25* aims to provide a Gold standard for CEA and a silver one for CTA and CPA. It therefore contains these kinds of annotations. CEA annotations are derived from explicit links in Wikipedia tables using the wikitextparser[17] library by checking the 'wikilinks' attribute of each table cell. The annotations within *MammoTab 25* derive from Wikidata v.20240720 and follow the structure used in the SemTab challenge. In this format, the CEA task expects each line to contain four fields: *Table ID*, *Row ID*, *Column ID*, and *Entity URI*. The CPA task requires each line to include the *Table ID*, *head column ID*, *tail column ID*, and the corresponding *Property URI*. The CTA task uses three fields: *Table ID*, *Column ID*, and *Annotation URI*.

[16] MammoTab documentation - unimib-datai.github.io/mammotab-docs/docs/introduction.

[17] wikitextparser - pypi.org/project/wikitextparser/.

Each table is stored as a CSV file, with each line representing a row. Annotations for CEA, CPA, and CTA are stored separately in dedicated CSVs. Additionally, a JSON file is generated for each source Wikipedia page, providing metadata and contextual features about the extracted tables (see Listing 1.1).

Listing 1.1. JSON document with the information relating to a Wikipedia page that contains at least one table.

```
{"wiki_id": ‘36115735’,
 "title": ‘As Long as You Love Me (Justin Bieber song)’,
 "tables": {
  "XXI7BFMW": {
    "caption": ‘Promotional release dates for "As Long as You Love Me"’,
    "header": [[‘Region’, ‘Date’, ‘Format’, ‘Label’]],
    "external_context": {},
    "link": [[‘’, ‘’, ‘’, ‘’],
             [‘’, ‘’, ‘Music_download’, ‘Island_Records’],
             ...],
    "text": [[‘Region’, ‘Date’, ‘Format’, ‘Label’],
             [‘United States’, ‘June 11, 2012’, ‘Digital Download’, ...],
             [‘Canada’, ‘June 11, 2012’, ‘Digital Download’, ...],
             ...],
    "target_col": [2],
    "entity": [[‘’, ‘’, ‘’, ‘’],
               [‘’, ‘’, ‘Q6473564’, ‘Q190585’],
               [‘’, ‘’, ‘Q6473564’, ‘Q190585’],
              ...],
    "types": [[[], [], [], []],
              [[], [], [‘Q81941037’], [‘Q18127’]]
             ...],
    "tags": {"0": {"col_type": "NE"},...},
    "col_types": [[], [], [[‘Q81941037’, 0.8571428571428571]], [[...
    "col_type_perfect": [‘’, ‘’, ‘Q81941037’, ‘’]}}},
    "stats": {
        "tot_linked_cell": 38,
        "entities_found": 37,
        "entities_not_found": 0,
        "types_found": 37,
        "types_not_found": 0,
        "tot_cells": 112,
        "nils": 1,
        "count_with_header": 1,
        "count_with_caption": 0,
        "acro_added": 1,
        "typos_added": 42,
        "approx_added": 52,
        "alias_added": 8,
        "generic_types": 26,
        "specific_types": 290,
        "filtered_types": 0,
        "found_perfect_types": 3,
        "tot_cols_with_types": 3,
        "count_single_domain": 0,
        "count_multi_domain": 1}}}
```

MammoTab 25 includes a rich set of contextual signals extracted from the original Wikipedia source page. These include the table's *caption*, *header*, *page title*, and associated *section titles and texts*, collectively stored in the `external_context` field. The `tags` field indicates the semantic type of each column, distinguishing between *Named Entities (NE)* and *Literals (LIT)*. Literal columns are further classified into granular types such as `WORD`, `NUMBER`, `QUOTED_PHRASE`, `URL`, `EMAIL`, `MENTION`, `HASHTAG`, `EMOJI`, `TIME`, `DATE`, `CURRENCY`, `PUNCTUATION`, and `SYMBOL`.

The `stats` field summarises each table's key statistics and structural attributes, supporting the creation of tailored evaluation subsets. This comprehensive representation is critical for enhancing semantic table interpretation by enabling models to utilise structural and document-level context. Notably, annotation without external context is significantly more challenging, as it requires

Table 2. The table lists the key challenges of the STI. For each challenge, the table specifies how it is addressed, which table element it affects, the tag used within the dataset to identify a subset of tables, and the number of tables that cover that specific challenge. Challenges marked with an asterisk (*) are not yet supported and will be incorporated in the next version of the dataset.

Challenge	Specs	Element	Tag	Number
1 Distinguishing between column types	NE-columns and L-columns	NE/L	—	4 261 797
	all NE-columns	NE	named	3 147 570
	all L-columns	L	literal	2 034 383
2 Heterogeneity of domains	single domain	NE	single domain	822 402
	multiple domains	NE	multi domains	16 528
3 Diversity of data	general	CELL	general domain	54 483 081
	specific	CELL	specific domain	236 657 820
4 Limited contextual information	with header	Table	header	767 644
	with caption	Table	caption	88 826
	with external context	Table	context	configurable
5 Matching tabular values against the KG	with acronyms	NE	acronym	302 903
	presence of aliases	NE	aliases	1 139 367
	presence of typos	NE	typos	1 924 917
	approximate values	L	approx	1 164 863
6 Disambiguation of named-entities*	presence of homonyms	NE	—	—
7 Identification of NIL-mentions	presence of NIL	NE	nil	4 808 759
8 Choosing the most appropriate types	generic types	NE	multi types	54 483 081
	specific types	NE	multi types	236 657 820
9 Choosing the most appropriate properties*	generic properties	NE/L	—	—
	specific properties	NE/L	—	—
10 Amount and shape of data	#rows - #columns	Table	—	838 930

models to deduce semantics based solely on intrinsic table features. Consequently, many benchmarks and evaluations deliberately target these low-context scenarios.

Table 2 shows for each key challenge how it is represented in the tables, which element of the table it affects (*i.e.*, column types or the entire table), the tag used in the metadata, and the number of tables within the dataset.

For example, the dataset includes tables that contain: i) entities that pose challenges in terms of disambiguation (*e.g.*, table id LBJJ1WGD - reactor Clinton)[18], ii) instances of homonymy (*e.g.*, table id MRBWAAOA - soccer player Michael Jordan)[19], iii) aliases (*e.g.*, table id JCND1XGG - Tom Riddle alias of Lord Voldemort)[20], and iv) NIL-mentions (*e.g.*, table id MSTBGKPR - KKOP-LP Wildcat Broad. Inc)[21].

[18] List of canceled nuclear reactors in the United States - en.wikipedia.org/wiki/List_of_cancelled_nuclear_reactors_in_the_United_States.

[19] USA Today All-USA High School Football Team - en.wikipedia.org/wiki /US_Today_All-USA_high_school_football_team.

[20] Christian Coulson - en.wikipedia.org/wiki/Christian_Coulson.

[21] List of radio stations in Nebraska - en.wikipedia.org/wiki/List_of_radio_stations_in_Nebraska.

MammoTab 25 has been published on Zenodo[22] following the FAIR Guiding Principles [29], under the GNU Affero General Public License v3.0 license[23].

4.1 Dataset Profile

The *MammoTab 25* tables were extracted from a staggering 63 million English Wikipedia pages using the XML dump. Out of these pages, ∼1 438 841 tables were detected. From this extensive selection, we identified and stored the tables that contained at least three links in the same column, resulting in a total of 838930 tables. Some dataset statistics are reported in Table 3.

Table 3. Overall statistics of the *MammoTab 25* dataset: the total number of *Tables*, *Columns*, and *Rows*; the minimum, maximum, and the average number of columns and rows per table. *Linked Cells* refers to the number of cells linked to an entity, *Typed Cols* refers to the number of columns with a known class associated, and *NILs* refers to the number of NIL-mentions.

Tables	Columns			Rows			Linked Cells	Typed Cols	NILs
	total	min	max	total	min	max			
838930	4 261 797	1	676	21 871 164	3	9 966	2 642 709	1 596 854	4 808 759

4.2 Implementation

The dataset was constructed using a pipeline (Fig. 4) comprised of a series of Python scripts. These scripts can be found in a Git repository[24]. The dataset is accompanied by documentation describing how to run and modify the pipeline[25].

The process begins with an initial configuration step, during which environment variables are set to customise the desired output[26].

The generation process includes several configurable parameters that allow fine-grained control over the characteristics of the synthesised tables. These parameters specify whether to introduce *acronyms*, *aliases*, *typographical errors*, and *approximate numerical values*, along with the frequency at which each alteration should occur.

Additional settings define structural constraints on the generated tables, such as the maximum number of *rows*, *columns*, and *header lines*. Moreover, users can choose whether to include *external context* in the exported dataset.

The pipeline comprises 12 steps:

[22] MammoTab 25 Zenodo page - doi.org/10.5281/zenodo.16562700.

[23] GNU Affero General Public License - www.gnu.org/licenses/agpl-3.0.html.

[24] GIT repository - github.com/unimib-datAI/mammotab.

[25] MammoTab Docs - unimib-datai.github.io/mammotab-docs/.

[26] MammoTab Docs - unimib-datai.github.io/mammotab-docs/docs/introduction.

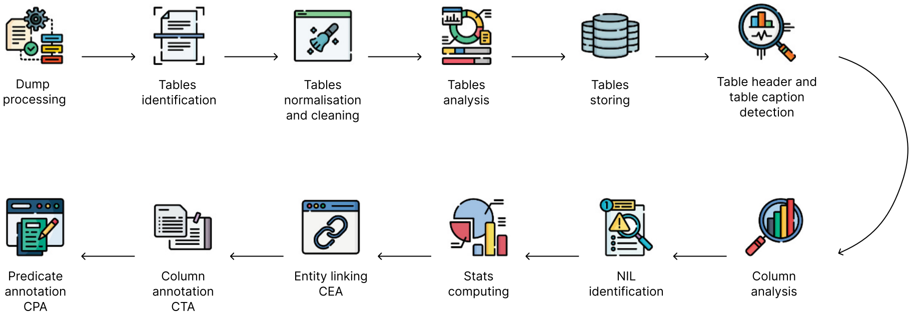


Fig. 4. Graphical representation of the *MammoTab 25* pipeline.

1. *dump processing*: each Wikipedia XML dump file (we used the multiple bz2 stream dumps for easier parallelisation) is parsed using WikiTextParser and BeautifulSoup[27] to extract each page;
2. *table identification*: pages are scanned to find those that contain at least one table (wikitext syntax: | class=wikitable |);
3. *table normalisation and cleaning*: each cell is normalised and cleaned up using custom rules[28] and wikitextparser[29] functions. Elements such as subscripts, superscripts, elements of the wikitext syntax, images, Wikipedia help and project pages links, and links to external pages are removed;
4. *table analysis*: each table is analysed to check for cells that contain links to Wikipedia pages (wikitext syntax: [[link]]). A cell is considered a mention of an entity only if the entire cell is a link. The remaining cells, which contain multiple links or additional words surrounding the link, may also be mentions of entities; however, we consider them uncertain and mark them as UNKNOWN. Future versions of the dataset will also consider these cases, provided that a reliable method is established to handle cells containing multiple links or additional surrounding text correctly;
5. *table storing*: tables that have at least three fully linked cells in a column are kept for export;
6. *table header and table caption detection*: table header (wikitext syntax: !header) and table caption, if any, are stored and added to the current table;
7. *column analysis*: each column is analysed and classified into Literal columns (L-column) for datatype values (*e.g.*, strings, numbers, dates, such as 4808, 10/04/1983), or into Named-Entity columns (NE-column) if it contains links to Wikipedia pages using the method described in [5] based on [32];
8. *NIL-identification*: we mark as NIL the cells containing Wikipedia red links[30], which are those links referring to a page that does not exist;

[27] BeautifulSoup - www.crummy.com/software/BeautifulSoup/.

[28] MammoTab Datacleaning techniques - unimib-datai.github.io/mammotab-docs/docs/data-cleaning.

[29] wikitextparser - github.com/5j9/wikitextparser.

[30] Wikipedia:Red link - en.wikipedia.org/wiki/Wikipedia:Red_link.

9. *stats computation*: per table statistics are computed;
10. *entity linking - CEA*: for each Wikipedia link, the related Wikidata entity is extracted (using *LamAPI* with the /entity/sameas endpoint, which allows mapping between Wikipedia pages and DBpedia or Wikidata entities);
11. *column annotation - CTA*: column types are set by choosing the most specific entity class (according to Wikidata subclass relationships) shared by most column rows. Like in [19] and similarly to [9], for columns with fewer than 5 rows, all cells must be instances of that class, whereas for bigger columns, at least 60% of the cells must be instances of that class, else the tables are not considered for CTA;
12. *predicate annotation - CPA*: predicates between two columns are identified using *LamAPI*with the /entity/predicates endpoint by passing each pair of entities between two columns, and the most common predicate is selected.

Since the CTA and CPA labels were generated through heuristic methods and benchmarked against state-of-the-art approaches, the associated MammoTab annotations should be considered a silver standard. To enhance dataset generation, we leveraged *LamAPI*, whose entity-retrieval layer, validated on T2Dv2 [24], Tough Tables [6], SemTab 2019 and SemTab2021, achieves about 90% coverage with only 20–30 candidates per mention [3], offering a markedly simpler and more effective alternative to raw SPARQL queries while still allowing the pipeline to fall back on direct SPARQL endpoints when needed.

The Listing 1.1 shows an example of a JSON document used to manage the results of the process described above on the Wikipedia page of "As Long as You Love Me (Justin Bieber song)"[31].

Re-running the Python scripts, simply pointing to a new Wikipedia XML dump file to process[32], enables generating an updated version of the dataset whenever the corresponding Wikipedia pages undergo some changes.

5 Evaluation

The experiments in this Section are intended to showcase how the use of *MammoTab 25* facilitates identifying weaknesses in an STI approach, explicitly focusing on addressing the primary challenges outlined in Sect. 3.

A subset of the *MammoTab 25* dataset was extracted by oversampling specific features to create a challenge-wise balanced evaluation set, ensuring representation across STI challenges. The resulting sample comprises 870 tables containing a total of 84,907 cells. This dataset will be employed as the primary test collection for the first round of the SemTab 2025 challenge. Table 4 provides detailed statistics for this evaluation dataset.

[31] As Long as You Love Me (Justin Bieber song) - en.wikipedia.org/wiki /As_Long _as_You_Love_Me_(Justin_Bieber_song).

[32] Wikimedia Downloads - dumps.wikimedia.org/backup-index.html.

Table 4. Sample dataset statistics

Tables	Cells		Types		Acronyms	Typos	Aliases	Domains	Rows			Columns		
	total	NIL	Generic	Specific					min	max	avg	min	max	avg
870	84 907	14 856	266 706	1 125 199	3 518	12 135	7 117	435	4	253	43,47	1	36	6,04

Nineteen Large Language Models (LLMs) were selected to populate a public leaderboard[33]. In addition, we evaluated TableLlama, the first autoregressive LLM explicitly instruction-tuned for tabular data, which currently delivers SOTA performance [31]. The results are presented in Figs. 5 and 6, which illustrate the overall and per-challenge accuracy, respectively. Open-source models were chosen because they can be tested freely and uniformly across. In contrast, many leading closed-source systems impose licensing restrictions and usage fees that would have made large-scale, side-by-side benchmarking impractical in an academic setting.

Each model was evaluated once using a greedy decoding strategy, *i.e.*, at each generation step, the token with the highest probability was selected. Generation was capped at 128 tokens, and the Key-Value cache [17] was enabled by default for all models. The temperature, also known as τ, was set to 0.7 for all models.

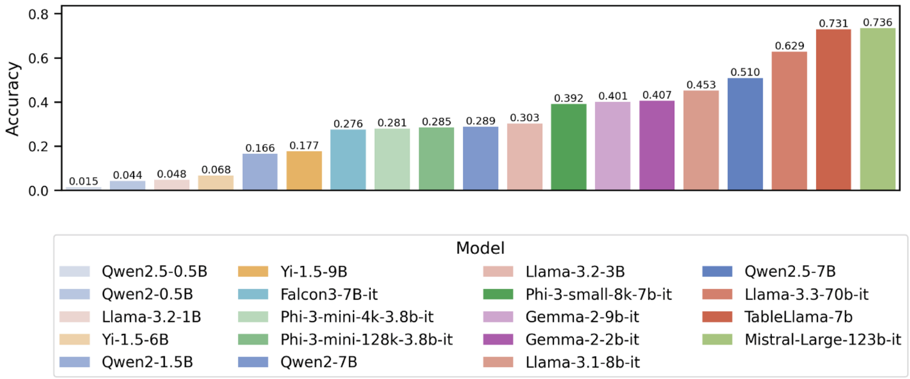


Fig. 5. Overall accuracy of the 19 evaluated LLMs on the *MammoTab 25* benchmark.

It emerges from Figs. 5 and 6 that larger models achieve substantially higher accuracy across all challenge categories. In addition to scale, specialised fine-tuning for tabular data (TableLlama-7b) confers a notable advantage, allowing mid-sized models to approach or match the performance of much larger general-purpose models (Mistral-Large-123b-it). Smaller models show markedly lower performance, especially on more complex or multi-domain tasks. All models struggle with the *NILs* challenge, achieving low accuracy (with a maximum of

[33] Mammotab Leaderboard - unimib-datai.github.io/mammotab-docs/leaderboard.

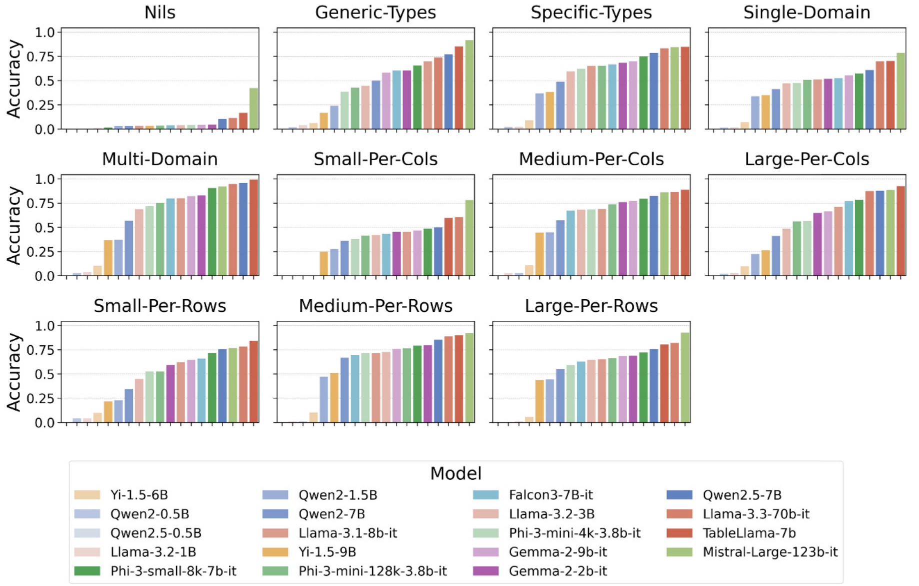


Fig. 6. Per-challenge accuracy of the 19 benchmarked LLMs.

47% achieved by Mistral-Large-123b-it), regardless of size, indicating persistent difficulty for current architectures. In contrast, accuracy is higher on data representing the *Multi-Domain* and *Large-Per-Rows* challenges, especially for the largest models with increased context sizes. *Small-Per-Cols* and *Small-Per-Rows* are harder than those with broader patterns, highlighting the need for more substantial inductive biases or better context aggregation. The relative ranking of models remains stable across challenge types: models that are strong overall tend to be strong in each category.

The findings indicate that the current version of *MammoTab 25* tables is a valuable asset for testing STI approaches, which must exhibit advanced techniques, considering various semantic aspects, by relating the performance of models with different characteristics to specific challenges. It is worth noting that, as seen in other datasets [6], it is possible to introduce some noise (*i.e.*, adding misspelt or fake mentions) to enhance the complexity of the annotation task, as depicted above.

The public leaderboard will be continuously updated as new language models are released, making the benchmark a living resource for the community.

6 Conclusion and Future Work

STI maps table elements (cells/mentions, columns, rows) to semantic tags (entities, classes, properties) from KG like Wikidata or DBpedia. The process includes tasks such as cell/mention to KG entity matching (CEA), column to KG class matching (CTA), and column-pair to KG property matching (CPA).

Despite the availability of several datasets for STI, it is recognised that these often cover only a subset of the characteristics of web tables and lack interpretability. In response to this gap, *MammoTab 25* was created. It comprises 838930 tables extracted from English Wikipedia pages and annotated through Wikidata. A preliminary release (v2-alpha) powered Round 2 of the "LLM vs STI" track at SemTab 2024, where three systems used it.

Researchers can use *MammoTab 25* to train, test and improve STI techniques, especially those reliant on large datasets. The dataset's vast size and diverse features make it a valuable resource for advancing research in this field.

A sample of *MammoTab 25* has been used in SemTab 2025, resulting in extensive dataset evaluation by various approaches and vice versa.

MammoTab 25 already covers eight of the ten challenges for STI, yet two starred challenges, namely Challenge 6 (Disambiguation of named-entities) and Challenge 9 (Choosing the most appropriate properties), remain unsupported. These gaps will be closed in the next iteration by adding the homonyms for the entities and the properties hierarchy for column-pair annotation. Looking forward, our development strategy also calls for continuously updating the corpus, enriching it with partial-text annotations. To make this maintenance plan robust rather than aspirational, an automated job will monitor every new English Wikipedia dump, trigger the full regeneration pipeline, execute integrity and consistency tests, and, provided all thresholds are met, automatically tag a release candidate.
Resource Availability Statement: Source code for generating MammoTab is available from GitHub[34]. The *MammoTab 25* dataset is available from Zenodo[35].

Acknowledgement. The work has been partly supported by the European Union's Horizon Europe research and innovation programme through the DataPACT (GA No. 101189771) and enRichMyData (GA No. 101070284) projects, as well as by MIUR's PRIN programme through the Discount Quality for Responsible Data Science project (GA No. 202248FWFS).

References

1. Abdelmageed, N., Schindler, S., König-Ries, B.: Biodivtab: a table annotation benchmark based on biodiversity research data. In: SemTab@ ISWC, pp. 13–18 (2021)
2. Avogadro, R., Cremaschi, M., D'adda, F., De Paoli, F., Palmonari, M.: Lamapi: a comprehensive tool for string-based entity retrieval with type-base filters. In: 17th ISWC workshop on ontology matching (OM), p. Online (2022)
3. Avogadro, R., D'Adda, F., Cremaschi, M.: Feature/vector entity retrieval and disambiguation techniques to create a supervised and unsupervised semantic table interpretation approach. Knowl.-Based Syst. **304**, 112447 (2024)

[34] github.com/unimib-datAI/mammotab.
[35] doi.org/10.5281/zenodo.16562700.

4. Ciavotta, M., Cutrona, V., De Paoli, F., Nikolov, N., Palmonari, M., Roman, D.: Supporting semantic data enrichment at scale. In: Technologies and Applications for Big Data Value, pp. 19–39. Springer (2022)
5. Cremaschi, M., De Paoli, F., Rula, A., Spahiu, B.: A fully automated approach to a complete semantic table interpretation. Futur. Gener. Comput. Syst. **112**, 478–500 (2020)
6. Cutrona, V., Bianchi, F., Jimenez-Ruiz, E., Palmonari, M.: Tough tables: carefully evaluating entity linking for tabular data. In: The Semantic Web - ISWC 2020, pp. 328–343. Lecture Notes in Computer Science, Springer International Publishing (November 2020)
7. Cutrona, V., et al.: Results of semtab 2021. In: 20th International Semantic Web Conference, vol. 3103, pp. 1–12. CEUR Workshop Proceedings (March 2022)
8. Cutrona, V., Ciavotta, M., Paoli, F.D., Palmonari, M.: ASIA: a tool for assisted semantic interpretation and annotation of tabular data. In: Proceedings of the ISWC 2019 Satellite Tracks. CEUR Workshop Proceedings, vol. 2456, pp. 209–212. CEUR-WS.org (2019)
9. Deng, X., Sun, H., Lees, A., Wu, Y., Yu, C.: TURL: table understanding through representation learning. ACM SIGMOD Rec. **51**(1), 33–40 (2022)
10. Hassanzadeh, O., et al.: Results of semtab 2024. In: CEUR Workshop Proceedings. vol. 3889, pp. 1–11 (2024)
11. Hulsebos, M., Demiralp, Ç., Groth, P.: Gittables: a large-scale corpus of relational tables. Proc. ACM Manage. Data **1**(1), 1–17 (2023)
12. Jimenez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K., Cutrona, V.: Results of semtab 2020. CEUR Workshop Proc. **2775**, 1–8 (January 2020)
13. Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K.: Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems. In: The Semantic Web, pp. 514–530. Springer International Publishing, Cham (2020)
14. Jiomekong, A., Etoga, C., Foko, B., Tsague, V., Folefac, M., Kana, S., Sow, M.M., Camara, G.: A large scale corpus of food composition tables. CEUR-WS. org, SemTab (2022)
15. Kejriwal, M., Knoblock, C.A., Szekely, P.: Knowledge graphs: Fundamentals, techniques, and applications. MIT Press (2021)
16. Korini, K., Peeters, R., Bizer, C.: Sotab: The wdc schema. org table annotation benchmark. Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org (2022)
17. Li, H., et al.: A survey on large language model acceleration based on KV cache management. arXiv preprint arXiv:2412.19442 (2024)
18. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. Proc. VLDB Endow. **3**(1–2), 1338–1347 (2010)
19. Marzocchi, M., Cremaschi, M., Pozzi, R., Avogadro, R., Palmonari, M.: Mammotab: a giant and comprehensive dataset for semantic table interpretation. In: Proceedings of the SemTab2022 (2022)
20. Mazurek, I., Wiewel, B., Kruit, B.: Wikary: A dataset of n-ary wikipedia tables matched to qualified wikidata statements. CEUR-WS. org, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2022)
21. Neumaier, S., Umbrich, J., Parreira, J.X., Polleres, A.: Multi-level semantic labelling of numerical values. In: The Semantic Web – ISWC 2016, pp. 428–445. Springer International Publishing, Cham (2016)

22. Palmonari, M., Ciavotta, M., De Paoli, F., Košmerlj, A., Nikolov, N.: Ew-shopp project: Supporting event and weather-based data analytics and marketing along the shopper journey. In: Advances in Service-Oriented and Cloud Computing, pp. 187–191. Springer International Publishing, Cham (2020)
23. Porrini, R., Palmonari, M., Cruz, I.F.: Facet annotation using reference knowledge bases. In: Proceedings of the 2018 World Wide Web Conference, pp. 1215–1224. WWW '18, WWW, Republic and Canton of Geneva, CHE (2018)
24. Ritze, D., Bizer, C.: Matching web tables to dbpedia - a feature utility study. In: Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017, pp. 210–221. OpenProceedings, Konstanz (2017)
25. Singh, S., Aji, A.F., Tomar, G.S., Christodoulopoulos, C.: Redtable: A relation extraction dataset for knowledge extraction from web tables. In: 29th International Conference on Computational Linguistics, pp. 2319–2327 (2022)
26. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: Leveraging linked data to discover semantic relations within data sources. In: The Semantic Web – ISWC 2016, pp. 549–565. Springer International Publishing, Cham (2016)
27. Tallet, P.: Les Papyrus de la Mer Rouge I: Le Journal de Merer. Institut Francais D'Archeologie Orientale (2017)
28. Weikum, G., Dong, X.L., Razniewski, S., Suchanek, F.M.: Machine knowledge: creation and curation of comprehensive knowledge bases. Found. Trends Databases **10**(2–4), 108–490 (2021)
29. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016)
30. Zhang, S., Meij, E., Balog, K., Reinanda, R.: Novel entity discovery from web tables. In: Proceedings of The Web Conference 2020, pp. 1298–1308. WWW '20, Association for Computing Machinery, New York, NY, USA (2020)
31. Zhang, T., Yue, X., Li, Y., Sun, H.: Tablellama: towards open large generalist models for tables (2024)
32. Zhang, Z.: Effective and efficient semantic table interpretation using tableminer+. Semantic Web **8**(6), 921–957 (2017)