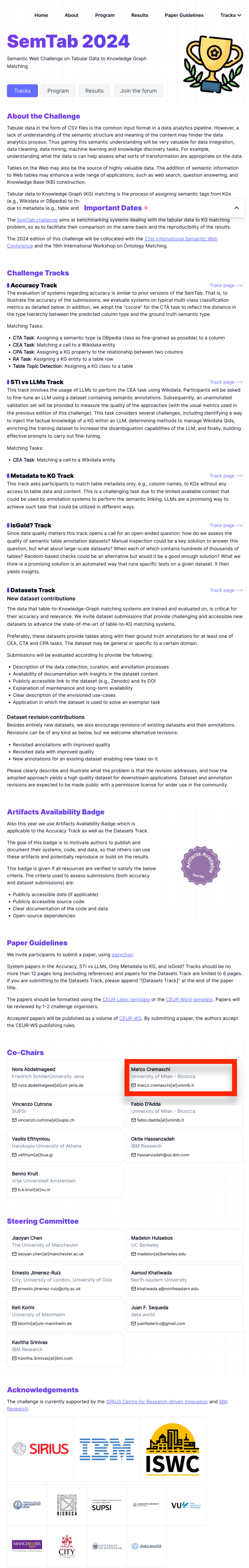

SemTab 2024

Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

About the Challenge

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However, a lack of understanding of the semantic structure and meaning of the content may hinder the data analytics process. Thus gaining this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what the data is can help assess what sorts of transformation are appropriate on the data.

Tables on the Web may also be the source of highly valuable data. The addition of semantic information to Web tables may enhance a wide range of applications, such as web search, question answering, and Knowledge Base (KB) construction.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from KGs (e.g., Wikidata or DBpedia) to th
due to metadata (e.g., table and

Important Dates

The SemTab challenge aims at benchmarking systems dealing with the tabular data to KG matching problem, so as to facilitate their comparison on the same basis and the reproducibility of the results.

The 2024 edition of this challenge will be collocated with the 23st International Semantic Web Conference and the 19th International Workshop on Ontology Matching.

Challenge Tracks

Accuracy Track

Track page ⟶

The evaluation of systems regarding accuracy is similar to prior versions of the SemTab. That is, to illustrate the accuracy of the submissions, we evaluate systems on typical multi-class classification metrics as detailed below. In addition, we adopt the "cscore" for the CTA task to reflect the distance in the type hierarchy between the predicted column type and the ground truth semantic type.

Matching Tasks:

- **CTA Task**: Assigning a semantic type (a DBpedia class as fine-grained as possible) to a column
- **CEA Task**: Matching a cell to a Wikidata entity
- **CPA Task**: Assigning a KG property to the relationship between two columns
- **RA Task**: Assigning a KG entity to a table row
- **Table Topic Detection**: Assigning a KG class to a table

STI vs LLMs Track

Track page ⟶

This track involves the usage of LLMs to perform the CEA task using Wikidata. Participants will be asked to fine-tune an LLM using a dataset containing semantic annotations. Subsequently, an unannotated validation set will be provided to measure the quality of the approaches (with the usual metrics used in the previous edition of this challenge). This task considers several challenges, including identifying a way to inject the factual knowledge of a KG within an LLM, determining methods to manage Wikidata Qids, enriching the training dataset to increase the disambiguation capabilities of the LLM, and finally, building effective prompts to carry out fine-tuning.

Matching Tasks:

- **CEA Task**: Matching a cell to a Wikidata entity

Metadata to KG Track

Track page ⟶

This track asks participants to match table metadata only, e.g., column names, to KGs without any access to table data and content. This is a challenging task due to the limited available context that could be used by annotation systems to perform the semantic linking. LLMs are a promising way to achieve such task that could be utilized in different ways.

IsGold? Track

Track page ⟶

Since data quality matters this track opens a call for an open-ended question: how do we assess the quality of semantic table annotation datasets? Manual inspection could be a key solution to answer this question, but what about large-scale datasets? When each of which contains hundreds of thousands of tables? Random-based checks could be an alternative but would it be a good enough solution? What we think is a promising solution is an automated way that runs specific tests on a given dataset. It then yields insights.

Datasets Track

Track page ⟶

New dataset contributions

The data that table-to-Knowledge-Graph matching systems are trained and evaluated on, is critical for their accuracy and relevance. We invite dataset submissions that provide challenging and accessible new datasets to advance the state-of-the-art of table-to-KG matching systems.

Preferably, these datasets provide tables along with their ground truth annotations for at least one of CEA, CTA and CPA tasks. The dataset may be general or specific to a certain domain.

Submissions will be evaluated according to provide the following:

- Description of the data collection, curation, and annotation processes
- Availability of documentation with insights in the dataset content
- Publicly accessible link to the dataset (e.g., Zenodo) and its DOI
- Explanation of maintenance and long-term availability
- Clear description of the envisioned use-cases
- Application in which the dataset is used to solve an exemplar task

Dataset revision contributions

Besides entirely new datasets, we also encourage revisions of existing datasets and their annotations. Revisions can be of any kind as below, but we welcome alternative revisions:

- Revisited annotations with improved quality
- Revisited data with improved quality
- New annotations for an existing dataset enabling new tasks on it

Please clearly describe and illustrate what the problem is that the revision addresses, and how the adopted approach yields a high quality dataset for downstream applications. Dataset and annotation revisions are expected to be made public with a permissive license for wider use in the community.

Artifacts Availability Badge

Also this year we use Artifacts Availability Badge which is applicable to the Accuracy Track as well as the Datasets Track.

The goal of this badge is to motivate authors to publish and document their systems, code, and data, so that others can use these artifacts and potentially reproduce or build on the results.

This badge is given if all resources are verified to satisfy the below criteria. The criteria used to assess submissions (both accuracy and dataset submissions) are:

- Publicly accessible data (if applicable)
- Publicly accessible source code
- Clear documentation of the code and data
- Open-source dependencies

Paper Guidelines

We invite participants to submit a paper, using easychair.

System papers in the Accuracy, STI vs LLMs, Only Metadata to KG, and isGold? Tracks should be no more than 12 pages long (excluding references) and papers for the Datasets Track are limited to 6 pages. If you are submitting to the Datasets Track, please append "[Datasets Track]" at the end of the paper title.

The papers should be formatted using the CEUR Latex template or the CEUR Word template. Papers will be reviewed by 1-2 challenge organisers.

Accepted papers will be published as a volume of CEUR-WS. By submitting a paper, the authors accept the CEUR-WS publishing rules.

Co-Chairs

Nora Abdelmageed
Friedrich SchillerUniversity Jena
✉ nora.abdelmageed[at]uni-jena.de

Marco Cremaschi
University of Milan - Bicocca
✉ marco.cremaschi[at]unimib.it

Vincenzo Cutrona
SUPSI
✉ vincenzo.cutrona[at]supsi.ch

Fabio D'Adda
University of Milan - Bicocca
✉ fabio.dadda[at]unimib.it

Vasilis Efthymiou
Harokopio University of Athens
✉ vefthym[at]hua.gr

Oktie Hassanzadeh
IBM Research
✉ hassanzadeh@us.ibm.com

Benno Kruit
Vrije Universiteit Amsterdam
✉ b.b.kruit[at]vu.nl

Steering Committee

Jiaoyan Chen
The University of Manchester
✉ jiaoyan.chen[at]manchester.ac.uk

Madelon Hulsebos
UC Berkeley
✉ madelon[at]berkeley.edu

Ernesto Jimenez-Ruiz
City, University of London, University of Oslo
✉ ernesto.jimenez-ruiz@city.ac.uk

Aamod Khatiwada
North-eastern University
✉ khatiwada.a@northeastern.edu

Keti Korini
University of Mannheim
✉ kkorini[at]uni-mannheim.de

Juan F. Sequeda
data.world
✉ juanfederico@gmail.com

Kavitha Srinivas
IBM Research
✉ Kavitha.Srinivas[at]ibm.com

Acknowledgements

The challenge is currently supported by the SIRIUS Centre for Research-driven Innovation and IBM Research.

SIRIUS
IBM
ISWC
Friedrich-Schiller-Universität Jena
Università degli Studi di Milano Bicocca
University of Applied Sciences and Arts of Southern Switzerland SUPSI
Harokopio University of Athens
VU Vrije Universiteit Amsterdam
Manchester 1824 The University of Manchester
City University of London Est 1894
University of Mannheim
data.world