

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 of the tables makes it difficult to automatically process them. This understanding will be very valuable for data integration, data cleaning, data mining, machine learning and many other applications. Therefore, the challenge is to understand what the data is and help assess what sorts of transformations 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 search engines, recommendation systems and data mining.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the table (e.g., rows, columns, cells) based on their metadata (e.g., table and column names) being missing, incomplete or ambiguous.

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 2021 edition of this challenge will be collocated with the 20th International Semantic Web Conference and the 18th International Workshop on Ontology Matching.

#### Datasets and Ground Truths

The ground truths are now open:

- Automatically Generated (AG) dataset (HardTables): [DOI](#) ([ISI](#)) ([archive](#)) ([dataset](#))
- HardTables (HT) dataset: [DOI](#) ([ISI](#)) ([archive](#)) ([dataset](#))
- BioTables dataset: [DOI](#) ([ISI](#)) ([archive](#)) ([dataset](#))
- BioTables dataset: [DOI](#) ([ISI](#)) ([archive](#)) ([dataset](#))
- GTTables dataset: [DOI](#) ([ISI](#)) ([archive](#)) ([dataset](#))

Target Knowledge Graphs: Schema.org (version: May 2021), DBpedia (version: 2018-10), Wikidata (version: 20210809)

The codes of the AIChallenge evaluators are also available [here](#).

#### SemTab @ ISWC 2021

See full ISWC program [here](#) with the relevant links to the sessions. Material from the SemTab sessions: [posters](#) and recorded oral presentations.

#### Results and Challenge Prizes

Results of all three rounds available [here](#). Summary of SemTab 2021 results [here](#).

Prizes sponsored by IBM Research:

- Applications track: BioTables dataset.
- Applications track: CEA-CPA dataset.
- Usability track: MTab team (1st Prize), JenTab (2nd Prize).

#### Papers

SemTab 2021 papers have been published in [Volume 31\(02\) CEUR-WS proceedings](#).

- [Roberto Albergo and Marco Cremonesi, \*MantisTable VI: A novel and efficient approach to Semantic Table Interpretation\*](#)

**Tabular Data Annotation with MTab Tool**  
 • Liqun Wang, Yingying Sheng, Jian Ding and Suhai Jin. **GMBTab: A Graph-Based Method for Interpreting Tabular Data via Knowledge Graph** (technical report).  
 • Bram Stuurwinkel, Filip De Turck and Femke Oostveld. **MAGIC: Mining an Augmented Graph using INK**, (technical report).  
 • Nora Abdellahmadi and Sirkis Schindler. **JenTab Meets SemTab 2021: New Challenges**.  
 • Nils Koenig, Philipp Hinsche, Stephan Stoll, Christian Geißler, and Michael Zettner. **OntoPipes: An Annotation Benchmark based on Biodiversity Research Data**.  
 • Yannick Bouvier, Manon Kadri, Sébastien Trelat, and Fabrice Chabot. **MTab: A Semantic Table Annotation Benchmark**.

• Yannick Bouvier, Manon Kadri, Sébastien Trelat, and Fabrice Chabot. **MTab: A Semantic Table Annotation Benchmark**.

• Yannick Bouvier, Manon Kadri, Sébastien Trelat, and Fabrice Chabot. **MTab: A Semantic Table Annotation Benchmark**.

• David Pichot, Sébastien Trelat, Fabrice Chabot, Frédéric Desch, Thomas Labé, Pierre Monnin and Raphaël Troncy. **DAOGTAB: Graph Contexts For Efficient Semantic Annotation Of Tabular Data**.

#### ISWC oral presentations

The results of the challenge will be presented on October 27 (Wednesday). Three teams will also present their systems.

October 27, Session 4D EDT (US): 10:20-11:20, CET (EU): 16:20-17:20, CST (China): 22:20-23:20:

- Challenge overview - 1 min. live. [[slides](#)]
- MTab (20min - 10 mins. recorded, [[video](#)])
- MTie - 10 min. recorded, [[video](#)]
- Announce awards, QA and wrap-up - 10 min. live. [[slides](#)]
- Announcement of awards, QA and wrap-up - 20 min. live. [[slides](#)]

#### ISWC poster presentations

SemTab will be present during the ISWC Posters & Demos/Social sessions. We will use [wonder.me](#) together with the other ISWC Semantic Web challenges.

Posters:

- SemTab summary
- MantisTable
- MTie
- DAOGTAB
- DAOGTABcode
- BioTables dataset
- MTab (in a different room as ISWC demo paper)

#### Ontology Matching workshop poster presentations

SemTab will also be present at the Ontology Matching (OM) workshop on October 25 (14:30-15:30 CET). See full OM program [here](#). We will also use [wonder.me](#) for the OM poster session (note that the wonder.me rooms are different).

#### Posters:

- JenTab
- MantisTable
- Magic

#### Participation: forum and registration

We have a [discussion group](#) for the challenge where we share the latest news with the participants and we discuss issues risen during the evaluation rounds.

Please register your system using this [google form](#).

Note that participants can join SemTab at any Round for any of the tasks/tracks.

#### Challenge Tasks

##### Accuracy Track

As in previous editions, SemTab includes the following tasks organised into several evaluation rounds:

- CTA Task: Assigning a semantic type to a DBpedia class as fine-grained as possible) to a column.
- CPA Task: Matching a DBpedia entity to a Wikidata entity.
- CTA-WD Task: Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column.

The challenge will be run with the support of the [AIChallenge](#) platform and the [STLTool](#) system.

##### Datasets and tasks per round

###### Round 1:

- Knowledge Graphs: Wikidata (version: 20210809)
- BioTable Datasets and targets: tables of CTA-DBP and CEA-CPA, CTA-DBP targets, CEA-CPA targets, tables of CTA-DBP and CEA-CPA targets, tables of BioTable-CTA-DBP and BioTable-CEA-CPA targets.
- CTA-DBP Task: Assigning a DBpedia semantic type (a DBpedia class as fine-grained as possible) to a column.
- CTA-WD Task: Matching a DBpedia entity to a Wikidata entity.
- CPA Task: Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column.
- CTA-WD Task: Matching a cell to a DBpedia entity. See [Alcaword page](#).
- CTA-DBP Task: Matching a cell to a DBpedia entity. See [Alcaword page](#).
- CTA-WD Task: Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column.
- CEA-CPA Task: Matching a cell to a Wikidata entity. See [Alcaword page](#).
- CEA-WD Task: Matching a cell to a Wikidata entity. See [Alcaword page](#).

###### Round 2:

- Knowledge Graphs: Wikidata (version: 20210809)
- BioTable Datasets and targets: tables of BioTable-CTA-DBP and BioTable-CEA-CPA, BioTable-CTA-DBP targets, BioTable-CEA-CPA targets, tables of BioTable-CTA-DBP and BioTable-CEA-CPA targets.
- BioTable-CTA-WD Task: Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column.
- BioTable-CEA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CPA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CTA-DBP Task: Matching a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CEA-CPA Task: Matching a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- HardTables R3 Datasets and targets: tables of HardTablesR3-CTA-DBP, HardTablesR3-CEA-CPA, HardTablesR3-CTA-WD, HardTablesR3-CEA-WD, HardTablesR3-CTA-CPA, HardTablesR3-CEA-CPA targets, HardTablesR3-CTA-WD targets, HardTablesR3-CEA-WD targets, HardTablesR3-CTA-CPA targets, HardTablesR3-CEA-CPA targets.
- HardTables-CEA-WD Task: Assigning a Wikidata entity to a cell. See [Alcaword page](#).
- HardTables-CPA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).

###### Round 3:

- Knowledge Graphs: Schema.org (version: May 2021), DBpedia (version: 2018-10), Wikidata (version: 20210809)
- BioTable Datasets and targets: tables of BioTable-CTA-DBP and BioTable-CEA-CPA, BioTable-CTA-DBP targets, BioTable-CEA-CPA targets, tables of BioTable-CTA-DBP and BioTable-CEA-CPA targets.
- BioTable-CTA-WD Task: Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column.
- BioTable-CEA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CPA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CTA-DBP Task: Matching a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- BioTable-CEA-CPA Task: Matching a Wikidata entity to a column pair (order matters). See [Alcaword page](#).
- HardTables R3 Datasets and targets: tables of HardTablesR3-CTA-DBP, HardTablesR3-CEA-CPA, HardTablesR3-CTA-WD, HardTablesR3-CEA-WD, HardTablesR3-CTA-CPA, HardTablesR3-CEA-CPA targets, HardTablesR3-CTA-WD targets, HardTablesR3-CEA-WD targets, HardTablesR3-CTA-CPA targets, HardTablesR3-CEA-CPA targets.
- HardTables-CEA-WD Task: Assigning a Wikidata entity to a cell. See [Alcaword page](#).
- HardTables-CPA-WD Task: Assigning a Wikidata entity to a column pair (order matters). See [Alcaword page](#).

###### Usability Track

This new track addresses a pain point in the community regarding the lack of publicly available easy-to-use generic solution that will address the needs of a variety of applications and settings. We will devise a clear scoring mechanism to rank every participant's solution in terms of several usability criteria as judged by a review panel, for example:

- 1. Is the solution open-source?
- 2. Does the solution require specific platform that could affect its use in common settings?
- 3. Does the solution require specific hardware and tuning for a new application/domain?
- 4. Is the solution offered as a public service?
- 5. Does the solution include a well-designed user interface?

###### Applications Track

This new track is aimed at addressing applications in real-world settings, that take advantage of the output of the matching system. Changing dataset proposals are also more than welcome.

**Bio-Track:** Due to advances in biological research techniques, massive data is constantly being produced in the biomedical domain and it is commonly published unstructured or tabular formats. This data is not trivial to integrate semantically due to the heterogeneity of the data and the lack of standardised vocabularies and ontologies. Specifically, for tabular data annotation, the representation of data can have a significant impact in performance since each entity can be represented by alphanumeric codes (e.g., chemical formulas or gene names) or even have multiple representations. In addition, the choice of the representation of entities, entity types, properties and relationships to existing datasets to speed-up the process of integrating new data in the domain.

###### Important Dates (tentative)

- April 26: First call for challenge participants.
- June 30 - July 31: Round 1.
- August 7 - September 1: Round 2.
- September 1: Submission of best participants in Rounds 1 and 2 are invited to present their results during the ISWC conference and the Ontology Matching workshop.
- October 20: System paper submissions (via [easyChair](#)).
- October 21: Application paper submissions (via [easyChair](#)).
- October 26-28: Challenge Presentation and prize announcement.
- November 15: Final version system papers (via [easyChair](#)).

#### System Papers

We encourage participants to submit a system paper using [easyChair](#). The paper should be no more than 12 pages long (excluding references) and formatted using the [LNCS Style](#). System papers will be reviewed by 1-2 challenge organizers.

Accepted system papers will be published as a volume of [CEUR-WS](#). By submitting a paper, the authors accept the [CEUR-WS publishing rules](#).

#### Organisation

This challenge is organized by [Kadirina Simicic](#) (IBM Research), [Bentua Jiménez-Ros](#) (City University of London), [University of Oxford](#), [Oxford Internet Institute](#) (University of Oxford), [Sirkis Schindler](#) (Friedrich Schiller University Jena, Germany). The tables provided in this challenge are based on [BioTables](#) ([DBpedia](#), [Wikidata](#)). These tables have been curated for the challenge. In the form provided here, they may be used for the challenge, only. Any publication on them must include a reference to contain citations of the underlying datasets. These citations will be made available after the challenge deadline.

#### Acknowledgements

The challenge is currently supported by the [SIRIUS Centre for Research-driven Innovation](#) and [IBM Research](#).

BioTables is created by [Nora Abdellahmadi](#), [Sirkis Schindler](#), [Birgit König-Rees](#), [Herr Norbert Chiar](#) for [Distributed Information Systems](#), [Friedrich Schiller University Jena](#), Germany. The tables provided in this challenge are based on [BioTables](#) ([DBpedia](#), [Wikidata](#)). These tables have been curated for the challenge. In the form provided here, they may be used for the challenge, only. Any publication on them must include a reference to contain citations of the underlying datasets. These citations will be made available after the challenge deadline.