

## SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

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 data may hinder the data analysis process. Thus gaining the 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 Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the tables. This task however is often difficult in practice due to 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 co-located with the **20th International Semantic Web Conference** and the **10th International Workshop on Ontology Matching**.



### Datasets and Ground Truths

The ground truths are now open:

- Automatically Generated (AG) dataset (HerdTables): [DOI: 10.26434/chemrxiv-2020-08-01](#)
- Tough Tables (TT) dataset: [DOI: 10.26434/chemrxiv-2020-08-01](#)
- BioTable dataset: [DOI: 10.26434/chemrxiv-2020-08-01](#)
- BioDiTab dataset: [DOI: 10.26434/chemrxiv-2020-08-01](#)
- GTTables dataset: [DOI: 10.26434/chemrxiv-2020-08-01](#)

Target Knowledge Graphs: Schema.org (version: May 2021), DBPedia (version: 2016-10), Wikidata (version: 20210828)

The codes of the AICrowd 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: [BioDiTab dataset](#).
- Accuracy track: [DAGOBAH Team](#).
- Usability track: [MTab Team](#) (1st Prize), [JenTab](#) (2nd Prize).

### Papers

SemTab 2021 papers have been published as [Volume 3103 CEUR-WS proceedings](#).

**Roberto Augado and Marco Cremaschi, MantisTable V: A novel and efficient approach to Semantic Table Interpretation**

- Tabular Data Annotation with MTab Tool**
- Liwenheng Yang, Shengwei Shen, Jingyi Ding and Jiahui Jin, GIMTab: A Graph-Based Method for Interpreting Noisy Semantic Table to Knowledge Graph (poster/abstract)**
- Samir Boudoukh, Filip De Troock and Fennis Chagnon, BACIC: Mining an Augmented Graph using NK, starting from a CSV**
- Nora Abdelmegied and Srijo Schröder, JenTab Meets SemTab 2021's New Challenges**
- Nora Abdelmegied, Sören Schröder and Birgitte König-Ries, BioDiTab: A Table Annotation Benchmark based on Biodiversity Research Data**
- Wiam Boudoukh, Marouen Kadroudi and Sami Faiz, Kepler-ABI at SemTab 2021**
- Yue-Pei Huang, Jialong Liu, Yuesi Chao, Frédéric Douze, Thomas Labbe, Pierre Morin and Raphaël Troncy, DAGOBAH: Table and 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 - 10 min. live. [\[slides\]](#)
- DAGOBAH - 10 min. recorded. [\[video\]](#)
- MTab - 10 min. recorded. [\[video\]](#)
- JenTab - 10 min. recorded. [\[video\]](#)
- 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.

- Oct 26, 18:30-19:00 CET
- Oct 27, 18:30-19:10 CET
- Oct 28, 19:00-19:30 CET

### Posters

- [SemTab summary](#)
- [MantisTable](#)
- [MMap](#)
- [DAGOBAH](#)
- [GIMTab/seedtable](#)
- [BioDiTab 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 offered).

### Posters

- [JenTab](#)
- [MantisTable](#)
- [MMap](#)

### 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/datasets.

### Challenge Tasks

#### Accuracy Track

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

- 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.
- CFA Task:** Assigning a KG property to the relationship between two columns.

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

#### Datasets and tasks per round

##### Round 1:

- Knowledge Graphs:** DBPedia (version: 2016-10) and Wikidata (version: 20210828)
- Datasets and targets:** tables of CTA-DBP and CEA-DBP, CTA-DBP targets, CEA-DBP targets, tables of CTA-WD and CFA-WD, CTA-WD targets, CFA-WD targets
- CTA-DBP Task:** Assigning a *Dbpedia* semantic type (a *Dbpedia* class as fine-grained as possible) to a column. See [Allowed page](#).
- CEA-DBP Task:** Matching a cell to a *Dbpedia* entity. See [Allowed page](#).
- CTA-WD Task:** Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column. See [Allowed page](#).
- CEA-WD Task:** Matching a cell to a Wikidata entity. See [Allowed page](#).

##### Round 2:

- Knowledge Graphs:** Wikidata (version: 20210828)
- BioTable Datasets and targets:** tables of BioDiTab-CTA-WD, BioDiTab-CEA-WD and BioDiTab-CFA-WD, BioDiTab-CTA-WD targets, BioDiTab-CEA-WD targets, BioDiTab-CFA-WD targets
- BioTable-CTA-WD Task:** Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column. See [Allowed page](#).
- BioTable-CEA-WD Task:** Assigning a Wikidata entity to a cell. See [Allowed page](#).
- BioTable-CFA-WD Task:** Assigning a Wikidata property to a column pair (order matters). See [Allowed page](#).
- AG (HerdTables) Datasets and targets:** tables of HerdTables-CTA-WD, HerdTables-CEA-WD and HerdTables-CFA-WD, HerdTables-CTA-WD targets, HerdTables-CEA-WD targets, HerdTables-CFA-WD targets
- HerdTable-CTA-WD Task:** Assigning a Wikidata semantic type (a Wikidata entity as fine-grained as possible) to a column. See [Allowed page](#).
- HerdTable-CEA-WD Task:** Assigning a Wikidata entity to a cell. See [Allowed page](#).
- HerdTable-CFA-WD Task:** Assigning a Wikidata property to a column pair (order matters). See [Allowed page](#).

##### Round 3:

- Knowledge Graphs:** Schema.org (version: May 2021), DBPedia (version: 2016-10), Wikidata (version: 20210828)
- BioDiTab Datasets and targets:** tables of BioDiTab-CTA-WD and BioDiTab-CEA-WD, BioDiTab-CTA-WD targets, BioDiTab-CEA-WD targets, Knowledge Graph: *Live Editor of Wikidata*
- GTTables Datasets and targets:** tables of GTTables-CTA-DBP and GTTables-CTA-SCH (one column by one Schema.org class such as schema:author and schema:URL), GTTables-CTA-DBP targets, GTTables-CTA-SCH labels (including *Dbpedia* properties), GTTables-CTA-SCH targets, GTTables-CTA-SCH labels (including properties and types from Schema.org)
- AG (HerdTables) R3 Datasets and targets:** tables of HerdTablesR3-CTA-WD, HerdTablesR3-CEA-WD and HerdTablesR3-CFA-WD targets, Knowledge Graph: Wikidata (version: 20210828)

#### Usability Track

This new track addresses a pain point in the community regarding a lack of publicly available easy-to-use and 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:

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

#### Applications Track

This new track aims at addressing applications in real-world settings that take advantage of the output of the matching systems. Challenging dataset proposals are also more than welcome.

**Bio-Tab:** Due to advances in biological research techniques, new data is constantly being produced in the biomedical domain and it is commonly published unstructured or in regular formats. This data is not fitted to integrate semantically due not only to its sheer amount but also to the complexity of the biological relations between entities. 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 synonyms. Therefore, the domain would greatly benefit from automated methods to map entities, entity types and properties 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 - 17:** September 7 - 17: Round 2.
- August/early September:** Best participants in Rounds 1 and 2 are invited to present their results during the ISWC conference and the Ontology Matching workshop.
- September 20 - October 16:** Round 3.
- October 20:** System paper submissions (via [easyChair](#)).
- October 25:** Ontology Matching workshop.
- October 28-29:** 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 [L<sup>A</sup>T<sub>E</sub>X](#) 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 organised by [Kavitha Srinivas](#) (IBM Research), [Ernesto Jiménez-Ruiz](#) (City, University of London; University of Oulu), [Otilia Heesman](#) (IBM Research), [Jinyan Chen](#) (University of Oxford), [Vasilis Illyiannidis](#) (FOCM - ICS), [Vincenzo Cutrona](#) (University of Milano - Bicocca), [Juan Sequera](#) (Data.world), [Dimitris Christos](#) (Universidade de Lisboa), [Celia Pesquita](#) (Universidade de Lisboa), [Nora Abdelmegied](#) (University of Jeddah), and [Mikolaj Ilkiewicz](#) (University of Amsterdam). If you have any problems/feedback with the datasets or any suggestions related to this challenge, do not hesitate to contact us via the [discussion group](#).

### Acknowledgements

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

BioDiTab is credited to Nora Abdelmegied, Sören Schröder, Birgitte König-Ries, Heinz Nadstorf Chair for Distributed Information Systems, Friedrich Schiller University Jena, Germany. The tables provided in this challenge are based on real biodiversity research datasets, but have been adapted for the challenge. In the form provided here, they may be used for the challenge, only any publication on challenge results needs to contain citations of the underlying datasets. These citations will be made available after the challenge deadline.