KEPLER-ASI at SemTab 2021

Wiem Baazouzi[1], Marouen Kachroudi[2], and Sami Faiz[3]

[1] Université de la Manouba, Ecole Nationale des sciences de l'informatique, Laboratoire de Recherche en génIe logiciel, Application Distribuées, Systèmes décisionnels et Imagerie intelligente, LR99ES26, Manouba 2010, Tunis, Tunisie.
wiem.baazouzi@ensi-uma.tn

[2] Université de Tunis El Manar, Faculté des Sciences de Tunis, Informatique Programmation Algorithmique et Heuristique, LR11ES14, 2092, Tunis, Tunisie
marouen.kachroudi@fst.rnu.tn

[3] Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, Laboratoire de Télédétection et Systèmes d'Information à Référence Spatiale, 99/UR/11-11, 2092, Tunis, Tunisie
sami.faiz@insat.rnu.tn

Abstract. In this paper, we present our system KEPLER-ASI, for the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab 2021). This system is participating for the second time in this campaign, bringing improvements and new technical aspects. KEPLER-ASI analyzes tabular data to be able to detect correct matches in Wikidata and DBPedia. It should be noted that each data resource, or each round of the campaign imposes a certain number of constraints, requiring advanced techniques. The aforementioned task turns out to be difficult for the machines, which requires an additional effort in order to deploy the cognitive capacity in the matching methods. KEPLER-ASI still relies on the SPARQL query to semantically annotate tables in Knowledge Graphs (KG), in order to solve the critical problems of matching tasks. The results obtained during the evaluation phase are encouraging and show the strengths of the proposed system.

Keywords: Tabular Data - Knowledge Graph - KEPLER-ASI - SPARQL

1 Introduction

It is evident that the World Wide Web encompasses and conveys very large volumes of textual information, in several forms: unstructured text, semi-structured model-based web pages (which represent data in the form widely recognized by key-value notation and lists). In this broad context, the methods aiming to extract information from these resources to convert them in a structured form have been the subject of several works [1, 2]. As an observation, it is evident that there is a lack of understanding of the semantic structure which can hamper the process of data analysis. This observation reveals a gap between data amounts.

Indeed, acquiring this semantic reconciliation will therefore be very useful for data integration, data cleansing, data mining, machine learning and knowledge discovery tasks. For example, understanding the data can help assess the appropriate types of transformation. Depending on the use and deployment scenario, tabular data are carefully conveyed to the Web in various formats. The majority of these datasets are available in tabular form (*e.g.*, CSV (Comma-Separated Values)). The main reason for the popularity of this format is its simplicity: many common office tools are available to facilitate their generation and use. Tables on the Web are a very valuable data source. Thus, injecting semantic information into arrays on the web has the potential to boost a wide range of applications, such as web searching, answering queries, and building Knowledge Bases (KB). Research reports that there are various issues with tabular data available on the Web, such as learning with limited labeled data, defining or updating ontologies, exploiting prior knowledge, and/or scaling up existing solutions. Therefore, this task is often difficult in practice, due to missing, incomplete or ambiguous metadata (*e.g.*, table and column names). In recent years, we have identified several works that can be mainly classified as supervised (in the form of annotated tables to carry out the learning task) [3–7] or unsupervised (tables whose data is not dedicated to learning) [8, 7]. To solve these problems, we propose a global approach named KEPLER-ASI, which addresses the challenge of matching tabular data to knowledge graphs.This method is based on previous work, which deals with ontology alignment [9–15].

This year's SemTab campaign differs from the last two sessions[4] [5], in that it deals with Wikidata and DBPedia. In this challenge, the input is a CSV file, but three different challenges had to be met :

1. **CTA :** A type of the Wikidata (or eventually DBPedia) ontology had to be assigned a class KG to a column (Column-Type Annotation).
2. **CEA :** A Wikidata or DBPedia entity had to be matched to the different cells (Cell-Entity Annotation).
3. **CPA :** A KG (Wikidata or DBPedia) property had to be assigned to the relationship between two columns (Columns Property Annotation).

Data annotation is a fundamental process in tabular data analysis [16, 17], it allows to infer the meaning of other information. Then deduce the meaning of tabular data relating to a Knowledge Graph. The data we used was based both on Wikidata and DBPedia. It should be noted that in a broader context, the data used and manipulated obey the triples format representation : subject ($\mathcal{S}$), a predicate ($\mathcal{P}$) and an object ($\mathcal{O}$). This notation ensures semantic navigability through the data and makes all data manipulation more fluid, explicit and reliable. Indeed, Cell Entity Annotation (CEA) matches a cell to a KG entity. At this level, we have to annotate each individual element of the subject ($\mathcal{S}$) and the object ($\mathcal{O}$). Column Property Annotation (CPA) assigns a KG property to

[4] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2019/
[5] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2020/

the relationship between two columns. The task is to find out which property of the two columns are connected to either Wikidata or DBPedia. Column Type Annotation (CTA) assigns connected semantic type to a column. Our goal is to design a fast and efficient approach to annotate tabular data with entities from Wikidata or DBPedia. Our approach combines a multitude of NLP and search and filter strategies, based on text preprocessing techniques. Experiments carried out in the context of SemTab 2021 for all tasks have shown encouraging results.

2 KEPLER-ASI approach

In this section, we will describe in detail the different stages of our system, while presenting some basic notions to highlight the technical issues identified.

2.1 Key notions

– ***Tabular Data*** : S is a two-dimensional tabular structure made up of an ordered set of N rows and M columns, as depicted by Figure 1. n_i is a row of the table (i = 1 ... N), m_j is a column of the table (j = 1 ... M). The intersection between a row n_i and a column m_j is $c_{i,j}$, which is a value of the cell $S_{i,j}$. The table contents can have different types (string, date, float, number, etc.).

- Target Table (S): M × N.
- Subject Cell: $S_{(i,0)}$ (i = 1, 2 ... N).
- Object Cell: $S_{(i,j)}$ (i = 1, 2 ... M),(j = 1, 2 ... N).

$$
\begin{array}{c}
 \\
Row_1 \\
 \\
 \\
Row_j \\
 \\
 \\
Row_M
\end{array}
\begin{array}{c}
\begin{array}{ccccc}
\mathrm{Col}_0 & & \mathrm{Col}_i & & \mathrm{Col}_N
\end{array} \\
\begin{pmatrix}
S_{1,0} & \dots & \dots & \dots & S_{1,N} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
S_{j,0} & \dots & S_{j,i} & \dots & S_{j,N} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
S_{M,0} & \dots & \dots & \dots & S_{M,N}
\end{pmatrix}
\end{array}
$$

Fig. 1. Target Table

– ***Knowledge Graph*** : Knowledge Graphs have been in the focus of research since 2012, resulting in a wide variety of published descriptions and definitions. The lack of a common core, a fact that is also indicated by Paulheim [18] in 2015. Paulheim listed in his survey of Knowledge Graph refinement, the minimum set of characteristics that must be present to distinguish Knowledge Graphs from other knowledge collections, which basically

restricts the term to any graph based knowledge representation. In the online reviewing [18], authors agreed that a more precise definition was hard to find at that point. This statement points out the need of a closer investigation and deeper reflection in this area. Farber and *al.* defined a Knowledge Graph as a Resource Description Framework (RDF) graph and stated that the term KG was coined by Google to describe any graph-based Knowledge Base (KB) [19]. Although this definition is the only formal one, it contradicts with more general definitions as it explicitly requires the RDF data model. In the following we present a detailed description of our contribution, namely KEPLER-ASI.

2.2 System description

In order to address the above mentioned SemTab challenge tasks, KEPLER-ASI is designed according to the workflow depicted by Figure 2. There are three major complementary modules which consist of, respectively, Preprocessing, Annotation context and Tabular data to KG matching. The aforementioned steps are the same for each round, but the changes remain minimal depending on the variations observed in each case.

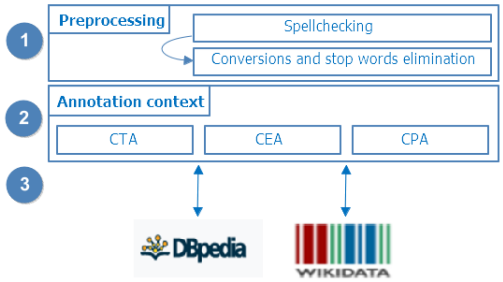

Fig. 2. Overview of our approach

As shown in Figure 2 Preprocessing aims to prepare the data inside the considered table. While Annotation Context, seeks to create a list of terms denoting the same context.

Preprocessing It should be noted that the content of each table can be expressed according to different types and formats, namely: numeric, character strings, binary data, date/time, boolean, addresses, etc. Indeed, with the great diversity of data types, the preprocessing step is crucial. Therefore, the goal of preprocessing is to ensure that the processing of each table is triggered without errors. The effort is especially accentuated when the data contain spelling errors. In other words, these issues must be resolved before we apply our approach. In order to well carry out this step, we used several techniques and libraries such as (Textblob[6], Pyspellchecker[7], etc.) to rectify and correct all the noisy textual

[6] https://textblob.readthedocs.io/en/dev/
[7] https://pypi.org/project/pyspellchecker/

data in the considered tables. As an example, we detect punctuation, parentheses, hyphen and apostrophe, and also stop words by using the `Pandas`[8] library to remove them. Like a classic treatment in this register, we ended this phase by transforming all the upper case letters into lower case.

Annotation context This phase allows to explicitly extract the candidates for the annotation process. The priming is carried out by an analysis of the processing columns, which aims to understand and delimit the set of regular expressions which contains a set of units: the area, the currency, the density, the electric current, the energy, flow rate, force, frequency, energy efficiency, unit of information, length, density, mass, numbers, population density, power, pressure, speed, temperature, time, torque, voltage and volume. This step allows to identify multiple Regextypes using regular expressions (*e.g.* numbers, geographic coordinates, address, code, color, URL). Since all values of type text are selected, preprocessing for natural languages was performed using the `langrid`[9] library to detect 26 languages in our data. By the way, it's a novelty for this year's SemTab campaign, *i.e.*, which makes the task more difficult with the introduction of natural language barriers. The `langrid` library is a stand-alone language identification tool. It is preformed on a large number of languages(97 currently). Doing so, correction, data type and language detection is performed. This can considerably reduce the effort and the cost of executing our approach by avoiding the massive repetition of these treatments for all the table cells, and this in each subtask.

Assigning a semantic type to a column (CTA) As depicted by Figure 3, the task is to annotate each entity column with elements from Wikidata (or possibly DBPedia) as its type identified during the preprocessing phase. Each item is marked with the tag in Wikidata or DBPedia. This treatment allows semantics identification. The CTA task can be performed based on Wikidata or DBPedia APIs which allows us to search for an item according to its description. The main information collected about a given entity and used in our approach are: a list of instances (expressed by the `instanceOf` primitive and accessible by the P31 code), the subclass of (expressed by the `subclassOf` primitive and accessible by code P279) and overlaps (expressed by the `partOf` primitive and accessible by code P361). At this point, we are able to process the CTA task using a SPARQL query. The SPARQL query is our interrogation mean fed from the main information of the entity which governs the choice of each data type, since they are a list of instances (P31), of subclasses (P279) or a part of a class (P361). The result of the SPARQL query may return a single type, but in some cases the result is more than one type, so in this case no annotation is produced for the CTA task.

[8] https://pandas.pydata.org
[9] https://github.com/openlangrid

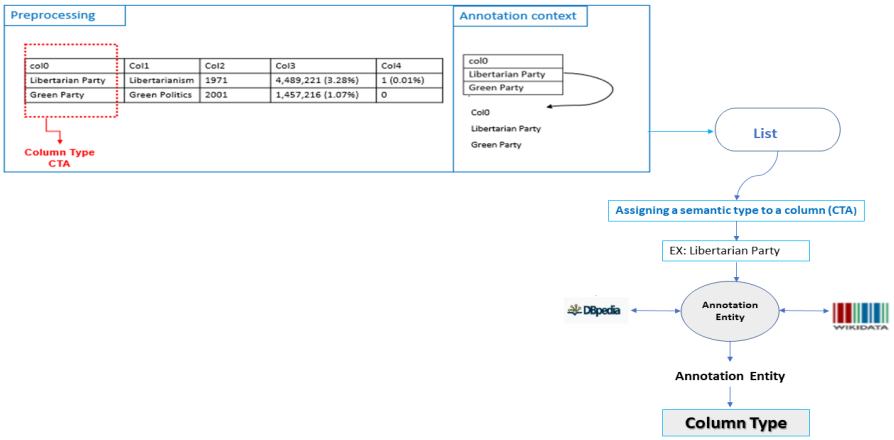

Fig. 3. CTA task at a glance.

Matching a cell to a KG entity (CEA) The CEA task aims to annotate the cells of a given table to a specific entity listed on Wikidata or DBPedia. Figure 4 gathers the CEA task that can be performed based on the same principle of CTA task. Our approach reuses the results of the CTA task process by introducing the necessary modifications on the SPARQL query. If the operation returns more than one annotation, we run a process based on examining the context of the considered column, relative to what was obtained with the CTA task, to overcome the ambiguity problem.

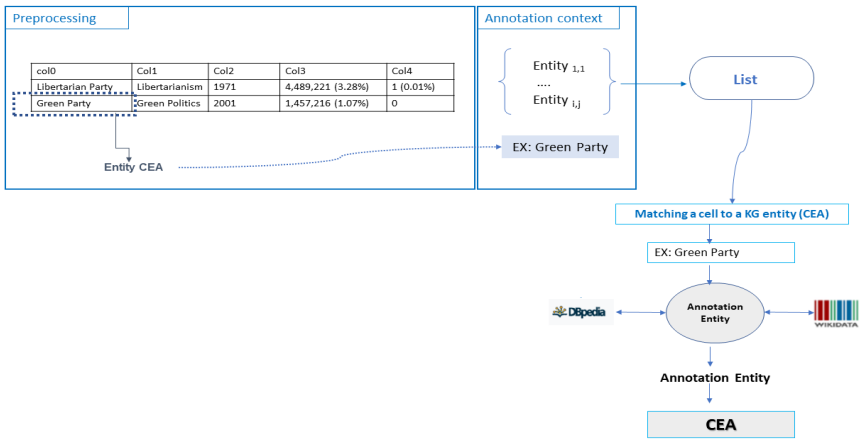

Fig. 4. Descriptive model of CEA task.

Matching a property to a KG entity (CPA) After having annotated the cell values as well as the different types of each of the considered entities, we will identify the relationships between two cells appearing on the same row via a property using a SPARQL query, as flagged by Figure 5. Indeed, the CPA task looks for annotating the relationship between two cells in a row via a property. Similarly, this latter task can be performed in an analogous manner to the CTA and CEA tasks. The only difference in the CPA task is that the SPARQL query must select both the entity and the corresponding attributes. The properties are fairly easy to match since we have already determined they during CEA and CTA task processing.

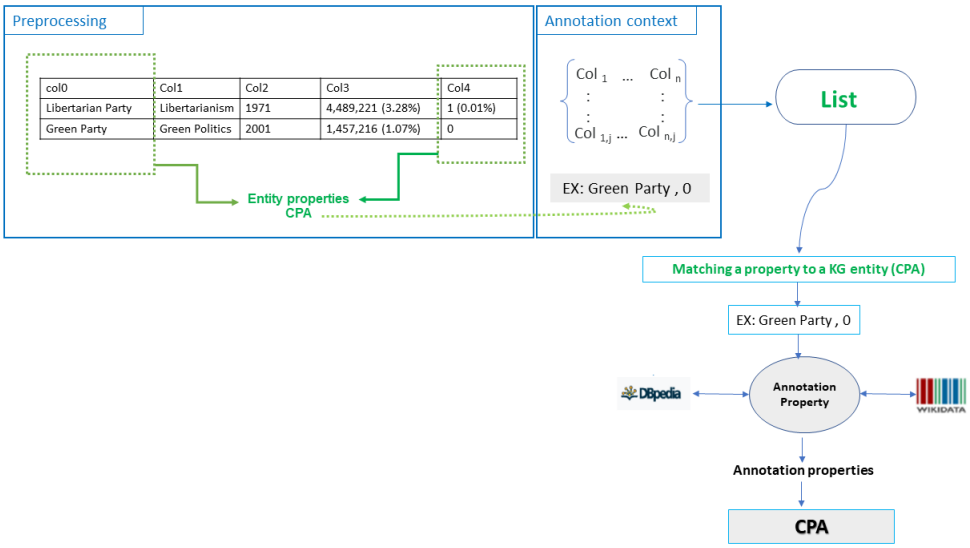

Fig. 5. A representation of CPA task.

3 KEPLER-ASI performance and results

In this section we will present the results of KEPLER-ASI for the different matching tasks in the 3 rounds of SemTab 2021. We would like to report that the results are presented according to two scenarios, *i.e.*, before deadline and after the deadline (since the organizers allow participants a period of 1 month before

freezing the values). Values are improved after the deadlines, as we finish the investigating work about the data specifics, thus adjusting our filters for the candidates identification. These results highlight the strengths of KEPLER-ASI with its encouraging performance despite the multiplicity of issues.

3.1 Round 1

In this first round, and in this version of SemTab 2021, four tasks are presented: CTA-WD, CEA-WD, CTA-DBP and CEA-DBP. The column type annotation (CTA -WD) assigns a Wikidata semantic type (a Wikidata entity) to a column. Cell Entity Annotation (CEA-WD) maps a cell to a KG entity. The processing carried out to search for correspondence on Wikidata is carried out in a similar way on Dbpedia.

Data for the CTA-WD and CEA-WD tasks were focused on Wikidata. As we explained in section 1, Wikidata is structured according to the RDF formalism, *i.e.*, subject (S), predicate (P) and Object (O). Each element considered is marked with a label in Wikidata, thus guaranteeing to take maximum advantage of its semantics. The CTA-WD and CEA-WD task data contains 180 tables. In Table 1, an example input table is provided. The first column contains an entity label, while the other columns contain the associated attributes.

Table 1. A data fragment for a table to match with Wikidata

Col0	**Col1**	**Col2**	**Col3**	**Col4**
Libertarian Party	Libertarianism	1971	4,489,221 (3.28%)	1 (0.01%)
Green Party	Green Politics	2001	1,457,216 (1.07%)	0

The column type annotation (CTA -DBP) assigns a DBPedia semantic type (a DBPedia entity) to a column. Cell Entity Annotation (CEA-DBP) matches a cell to an entity on the Knowledge Graph. The CTA-DBP and CEA-DBP task data also contains 180 tables. The results are summarized in Table 2.

Table 2. Results for Round 1

	F1 Score	**Precision**	**Rank**
CTA- WD	0.464	0.481	4
CTA-WD (after deadline)	0.746	0.758	3
CEA-WD	0.194	0.760	5
CEA-WD (after deadline)	0.620	0.841	3
CTA- DBP	0.027	0.133	5
CTA-DBP (after deadline)	0.391	0.520	3
CEA-DBP	0.110	0.644	5
CEA-DBP (after deadline)	0,509	0.610	4

In Round 1, we focused particularly on the preprocessing phase in order to choose and validate the spellchecker according to textual information, which can significantly improve the relative results of the CEA and CTA tasks. In summary, our review resulted in the use of two correctors, namely, Texteblob and Pyspellchecker. Both of these tools are intuitive, easy to use, and perform well in terms of Natural Language Processing (NLP).

During Round 1, the data size factor was impacting. We recognize that this round highlights the limits of machines in the face of such information volumes. Therefore, we can conclude that faced with this situation, the computing power and the speed of access to the external resources representing the Knowledge Graphs (*i.e.*, Wikidata and DBPedia) are decisive. In addition, we consider that the introduction of the cross-lingual aspect of this campaign has accentuated the challenge and allowed us to approach real scenarios that open and unlock the eventualities of the different proposed approaches applicability. Indeed, to support the cross-lingual aspect we acted at the level of the SPARQL query, as indicated on the code listing 1.1 , to automatically change the language label, and collect the candidates in any language. Thus, we have ensured the genericity of our SPARQL query, based on previous contributions [20, 15, 21].

```
1
2   endpoint_url = "############"
3
4   query = """
5   SELECT ?itemLabel ?class  ?property
6   WHERE {
7    ?item    ?itemDescription "%s"@en .
8    ?item wdt:P31 ?class
9      }
10   """
```

Code Listing 1.1. SPARQL query

3.2 Round 2

In Round 2, despite the distinction of the data and their grouping into two different families, they had a biological tint. Due to advances in biological research techniques, new data are constantly being generated in the biomedical field and they are routinely published in unstructured or tabular form. These data are not easy to integrate semantically, due not only to their size, but also to the complexity of the biological relationships maintained between the entities. Summary of metrics for this round is in Table 3.

Specifically, for tabular data annotation, the data representation can have a significant impact on performance since each entity can be represented by alphanumeric codes (*e.g.* chemical formulas or gene names) or even have multiple synonyms. Therefore, the studied field would greatly benefit from automated

Table 3. Results for Round 2

	F1 Score	**Precision**	**Rank**
BioTable-CTA- WD	0.811	0.811	6
BioTable-CTA- WD (after deadline)			
BioTable-CEA-WD	0.347	0.811	6
BioTable-CEA-WD (after deadline)	0.677	0.798	
BioTable-CPA-WD	0.853	0.880	4
BioTable-CPA-WD (after deadline)			
HardTable-CTA-WD	0.894	0.931	5
HardTable-CTA-WD (after deadline)			
HardTable-CEA-WD	0.707	0.919	6
HardTable-CEA-WD (after deadline)			
HardTable-CPA-WD	0.915	0.989	5
HardTable-CPA-WD (after deadline)			

methods to map entities, entity types, and properties to existing datasets to speed up the process of integrating new data into the domain. In this round the focus was on Wikidata, through two test cases: BioTable and HardTable. The different tasks: BioTable-CTA-WD, BioTable-CEA-WD and BioTable-CPA-WD on the one hand, to which we add Hard-CTA-WD, Hard-CEA-WD and Hard-CPA-WD, are all carried out on 110 tables.

During Round 2, we focused on the disambiguation problem. We have to decide when obtaining several candidates after querying the KGs. Indeed, our approach put in place during Round 1 was very useful and allowed us to reuse certain achievements. At this stage, we affirm that the automatic elements disambiguation processing remains a tedious task, given what it generates as an effort of semantic analysis and interpretation. Indeed, we have opted for the use of an external resource, namely Uniprot[10] [22]. UniProt integrates, interprets and standardizes data from multiple selected resources to add biological knowledge and associated metadata to protein records and acts as a central hub from which users can connect to 180 other resources. UniProt was recognized as an ELIXIR core data resource in 2017 [23] and received CoreTrustSeal certification in 2020. The data resource fully supports Findable, Accessible, Interoperable and Reusable, thus concretizing the (FAIR) data principles [24], for example by making data available in a number of community recognized formats, such as text, XML and RDF and through application programming interfaces (APIs) and FTP (File Transfer) downloads Protocol, providing traceable identifiers for protein sequences and protein sequence characteristics and fully highlighting data sources. The UniProt 2020 version contains over 189 million sequence records, with over 292,000 proteins, the complete set of proteins assumed to be expressed by an organism, derived from viral, bacterial, Archean and eukaryotic genomes

[10] https://www.uniprot.org

complete sequences available via UniProtKB Portail Proteomes[11]. In our case, Uniprot is used to support our disambiguation process. In other words, if there is a multiplicity of candidates in the matching process, or if there are no candidates, access to Uniprot allows us to overcome this problem.

Doing so, we end up with the scenario represented by the Figure 6. In fact, logically the processing of KEPLER-ASI ends at the stage, by obtaining the candidates likely to meet the need for matching. However, in some cases this answer may require some refinement.In case of multiple answers, Uniprot can help us to decide, given its richness and its ample description. In addition, in the absence of matching candidates (name differences, formulas, etc.), we can get the answer from Uniprot. Steps 4 and 5 are in addition to the regular KEPLER-ASI process, ensuring the redirection to Uniprot and the collection of any responses.

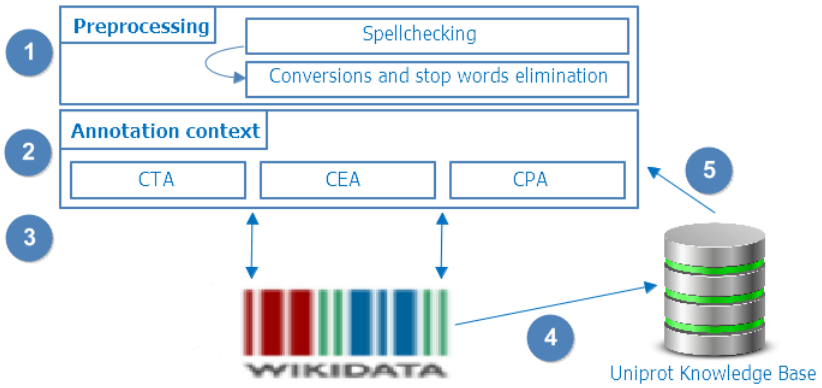

Fig. 6. Access to Uniprot and its contribution to KEPLER-ASI.

3.3 Round 3

Round 3 has 3 main test families:

- BioDiv: represented by 50 tables;
- GitTables: represented by 1100 tables;
- HardTables: represented by 7207 tables.

It should be noted that the stakes are the same for this round, moreover the evaluation is blind, *i.e.*, the participants do not have access to the evaluation

[11] https://www.uniprot.org/proteomes/

platform and its options. In other words, there is no test opportunity to adjust the parameters of the approach, according to the characteristics of the input. In this round too, we have opted for Uniprot to carry out treatments similar to those described in Round 2.

Table 4. Results for Round 3

	F1 Score	**Rank**
CEA-BioDiv	duplicate cells or columns	-
CTA-BioDiv	0.593	1
CEA-HARD	duplicate cells or columns	-
CTA-HARD	0.244	6
CPA-HARD	duplicate cells or columns	-
GIT-DBP	0.041	2
GIT-SCH	duplicate cells or columns	-

Out of the 7 proposed tasks, Kepler-aSI managed to process 3. In the CTA-BioDiv task, we are ranked first, for the GIT-DBP task we are ranked second and for CTA-HARD we are ranked sixth. For the other cases, our method produced outputs containing duplications, whereas these correspondences do not allow us to obtain evaluation metrics in order to be ranked.

4 Conclusion & Future Work

To summarize and conclude, we have presented in this paper the second version of our Kepler-aSI approach. Our system is participating in the challenge for the second time, it is approaching maturity and achieving very encouraging performance. We have succeeded in combining several strategies and treatment techniques, which is also the strength of our system. We boosted the preprocessing and spellchecking steps that got the system up and running.

In addition, despite the data size, which is quite large, we managed to get around this problem by using a kind of local dictionary, which allows us to reuse already existing matches. Thus, we realized a considerable saving of time, which allowed us to adjust and rectify after each execution. We also participated in all the tasks without exception, which allowed us to test our system on all facets, *i.e.*, to identify its strengths and weaknesses.

We tackled the several proposed tasks. Our solution is based on a generic SPARQL query using the cell contents as a description of a given item. In each round, despite the time allocated by the organizers running out, we continued the work and the improvements, having the conviction that each effort counts and brings us closer to the good control of the studied field.

KEPLER-ASI is a promising approach, but which will be further improved: First, we will apply several methods yet to correct spelling mistakes and other typos in the source data. Finally, we will try to develop our system by integrating new data processing techniques (some Big Data oriented paradigms). Indeed, the parallel implementation will allow us to circumvent the data size problem, which is the major gap for our current machines. Eventually, the idea of moving to a data representation using indexes [25, 26] would be a good track to investigate in order to master the search space, formed by the considered tabular data.

References

1. Chen, J., Jiménez-Ruiz, E., Horrocks, I., Sutton, C.: Colnet: Embedding the semantics of web tables for column type prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. Volume 33. (2019) 29–36
2. Malyshev, S., Krötzsch, M., González, L., Gonsior, J., Bielefeldt, A.: Getting the most out of wikidata: Semantic technology usage in wikipedia's knowledge graph. In: International Semantic Web Conference, Springer (2018) 376–394
3. Pham, M., Alse, S., Knoblock, C.A., Szekely, P.: Semantic labeling: a domain-independent approach. In: International Semantic Web Conference, Springer (2016) 446–462
4. Taheriyan, M., Knoblock, C.A., Szekely, P., Ambite, J.L.: Learning the semantics of structured data sources. Journal of Web Semantics **37** (2016) 152–169
5. Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P.: Assigning semantic labels to data sources. In: European Semantic Web Conference, Springer (2015) 403–417
6. Knoblock, C.A., Szekely, P., Ambite, J.L., Goel, A., Gupta, S., Lerman, K., Muslea, M., Taheriyan, M., Mallick, P.: Semi-automatically mapping structured sources into the semantic web. In: Extended Semantic Web Conference, Springer (2012) 375–390
7. Cremaschi, M., De Paoli, F., Rula, A., Spahiu, B.: A fully automated approach to a complete semantic table interpretation. Future Generation Computer Systems (2020)
8. Zhang, Z.: Effective and efficient semantic table interpretation using tableminer+. Semantic Web **8**(6) (2017) 921–957
9. Zghal, S., Kachroudi, M., Ben Yahia, S., Mephu Nguifo, E.: OACAS: Ontologies alignment using composition and aggregation of similarities. In: Proceedings of the 1st International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), Madeira, Portugal (2009) 233–238
10. Kachroudi, M., Ben Moussa, E., Zghal, S., Ben Yahia, S.: Ldoa results for oaei 2011. In: Proceedings of the 6th International Workshop on Ontology Matching (OM-2011) Colocated with the 10^{th} International Semantic Web Conference (ISWC-2011), Bonn, Germany (2011) 148–155
11. Kachroudi, M., Diallo, G., Ben Yahia, S.: OAEI 2017 results of KEPLER. In: Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21, 2017. Volume 2032 of CEUR Workshop Proceedings., CEUR-WS.org (2017) 138–145
12. Kachroudi, M., Ben Yahia, S.: Dealing with direct and indirect ontology alignment. J. Data Semant. **7**(4) (2018) 237–252

13. Kachroudi, M., Diallo, G., Ben Yahia, S.: KEPLER at OAEI 2018. In: Proceedings of the 13th International Workshop on Ontology Matching co-located with the 17th International Semantic Web Conference, OM@ISWC 2018, Monterey, CA, USA, October 8, 2018. Volume 2288 of CEUR Workshop Proceedings., CEUR-WS.org (2018) 173–178
14. Kachroudi, M., Zghal, S., Ben Yahia, S.: Bridging the multilingualism gap in ontology alignment. International Journal of Metadata, Semantics and Ontologies **9**(3) (2014) 252–262
15. Kachroudi, M., Zghal, S., Ben Yahia, S.: Using linguistic resource for cross-lingual ontology alignment. International Journal of Recent Contributions from Engineering **1**(1) (2013) 21–27
16. Chen, J., Jiménez-Ruiz, E., Horrocks, I., Sutton, C.: Learning semantic annotations for tabular data. arXiv preprint arXiv:1906.00781 (2019)
17. Efthymiou, V., Hassanzadeh, O., Rodriguez-Muro, M., Christophides, V.: Matching web tables with knowledge base entities: from entity lookups to entity embeddings. In: International Semantic Web Conference, Springer (2017) 260–277
18. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS) **48** (2016) 1–4
19. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. Semantic Web **9**(1) (2018) 77–129
20. Kachroudi, M., Ben Yahia, S., Zghal, S.: Damo - direct alignment for multilingual ontologies. In: Proceedings of the 3^{rd} International Conference on Knowledge Engineering and Ontology Development (KEOD), 26-29 October, Paris, France (2011) 110–117
21. Kachroudi, M., Zghal, S., Ben Yahia, S.: When external linguistic resource supports cross-lingual ontology alignment. In: Proceedings of the 5^{th} International Conference on Web and Information Technologies (ICWIT 2013), 9-12, May, Hammamet, Tunisia (2013) 327–336
22. Ruch, P., Teodoro, D., Consortium, U., et al.: Uniprot. Technical report (2021)
23. Drysdale, R., Cook, C.E., Petryszak, R., Baillie-Gerritsen, V., Barlow, M., Gasteiger, E., Gruhl, F., Haas, J., Lanfear, J., Lopez, R., et al.: The elixir core data resources: fundamental infrastructure for the life sciences. Bioinformatics (2020)
24. Garcia, L., Bolleman, J., Gehant, S., Redaschi, N., Martin, M.: Fair adoption, assessment and challenges at uniprot. Scientific data **6**(1) (2019) 1–4
25. Kachroudi, M., Diallo, G., Ben Yahia, S.: Initiating cross-lingual ontology alignment with information retrieval techniques. In: Actes de la $6^{\text{ème}}$ Edition des Journées sur les Ontologies (JFO'2016), Bordeaux, France (2016) 57–68
26. Zghal, S., Kachroudi, M., Damak, S.: Alignement d'ontologies base d'instances indexées. In: Actes de la $6^{\text{èmes}}$ Edition des Journées Francophones sur les Ontologies (JFO'2016), Bordeaux, France (2016) 69–74