DAGOBAH UI: A New Hope For Semantic Table Interpretation

Christophe Sarthou-Camy, Guillaume Jourdain, Yoan Chabot, Pierre Monnin, Frédéric Deuzé, Viet-Phi Huynh, Jixiong Liu, Thomas Labbé, Raphael Troncy

DAGOBAH UI: A New Hope For Semantic Table Interpretation

Christophe Sarthou-Camy[1], Guillaume Jourdain[1], Yoan Chabot[1], Pierre Monnin[1], Frédéric Deuzé[1], Viet-Phi Huynh[1], Jixiong Liu[1], Thomas Labbé[1], and Raphael Troncy[2]

[1] Orange, France
yoan.chabot@orange.com
[2] EURECOM, Sophia-Antipolis, France

Abstract. The past few years have seen a growing research interest in Semantic Table Interpretation (STI), i.e. the task of annotating tables with elements defined in knowledge graphs (KGs). These semantic annotations make use of entities and standardized types and relations and can, in turn, support several downstream use cases for tabular data such as dataset profiling and indexing, recommender systems, or dataset search. In this paper, we introduce DAGOBAH UI, a user-friendly Web interface that enables to visualize, validate, and manipulate results of STI methods such as the DAGOBAH API. Through an interactive demonstration on real world datasets, we illustrate how such a UI can ease the adoption and mass usage of STI techniques by end-users. A video of the demonstration is available at `https://tinyurl.com/dagobah-ui-demo`. An access to the DAGOBAH API can also be requested at `https://developer.orange.com/apis/table-annotation` (for logged in users).

Keywords: Semantic Table Interpretation · Table Enrichment · Knowledge Graph · DAGOBAH

1 Introduction

Large parts of available data either on the Web or in internal repositories of companies are encoded in tabular formats (e.g. Excel or CSV files) or as web tables [1]. Hence, there is a strong interest in understanding the semantics of such tables to pave the way for semantic-based services such as dataset search/recommendation, or enrichment of heterogeneous tabular datasets [2]. This process is known as Semantic Table Interpretation (STI) and has seen a growing research interest over the past few years, for example, through the SemTab challenge [6]. The 2021 edition of the SemTab challenge has featured a "Usability" track to foster research in user-friendly interfaces that will ease mass adoption of STI techniques. In this line of research, we propose DAGOBAH UI, a Web user interface that makes use of the RESTful API exposing the DAGOBAH SL system [5] – the best performing system at the SemTab 2021 Challenge – and the Wikidata KG [9].

2 Related Work

The closest tools to DAGOBAH UI are MantisTable [3] and the MTab interface [8]. MantisTable is a Web application that enables to import and manage tables as well as trigger specific annotation methods of the STI process. MTab is a Web interface that allows to upload a table and display the resulting semantic annotations provided by the MTab tool. An entity search functionality from a text input by the user is also provided. DAGOBAH UI offers additional functionalities such as the enrichment of tables with information from the knowledge graph, and conversely the enrichment of the knowledge graph with information from the tables. OpenRefine[3] is a powerful Web-based tool for cleaning, transforming, and extending tabular data with external data. The reconciliation functionalities are closed to the Cell-Entity Annotation task of the STI process and benefit now from a standardized protocol.[4] However, while a end-user has to manually instruct OpenRefine how to annotate specific columns via ad-hoc rules, the STI process is fully automatic in DAGOBAH UI.

3 System Description

DAGOBAH UI is a user-friendly Web interface that allows the manipulation, interpretation, and enrichment of relational tabular data. From a technical point of view, DAGOBAH is made of a NodeJS API using the Open API standard[5] and a frontend developed with VueJS and Boosted[6].

Tabular data files can be loaded in DAGOBAH UI from the local file system similarly to what OpenRefine and MantisTable [3] offer. In addition, DAGOBAH UI provides pre-loaded tabular data from the most commonly used benchmarks such as the SemTab datasets [6] and T2Dv2 [7].

3.1 Table Interpretation

DAGOBAH UI makes use of the RESTful API exposing DAGOBAH-SL [5], a system providing functionalities for table pre-processing and Semantic Table Interpretation. DAGOBAH-SL annotates tabular data with elements of KGs such as Wikidata or DBpedia. This system leverages *(i)* an Elasticsearch-based lookup service, and *(ii)* a disambiguation algorithm that relies on syntactic distances and comparisons between the context of an entity in the table and in the knowledge graph. This system was empirically evaluated during the three editions of the SemTab challenge [6] and has shown competitive performance with a 1st prize in the Accuracy track of the 2021 edition.

The preprocessing toolbox of DAGOBAH UI allows to clean a table (encoding problems, cell misalignment, etc.) as well as to extract information about the

[3] `https://openrefine.org/`
[4] `https://reconciliation-api.github.io/specs/latest/`
[5] `https://www.openapis.org/`
[6] `https://boosted.orange.com/`

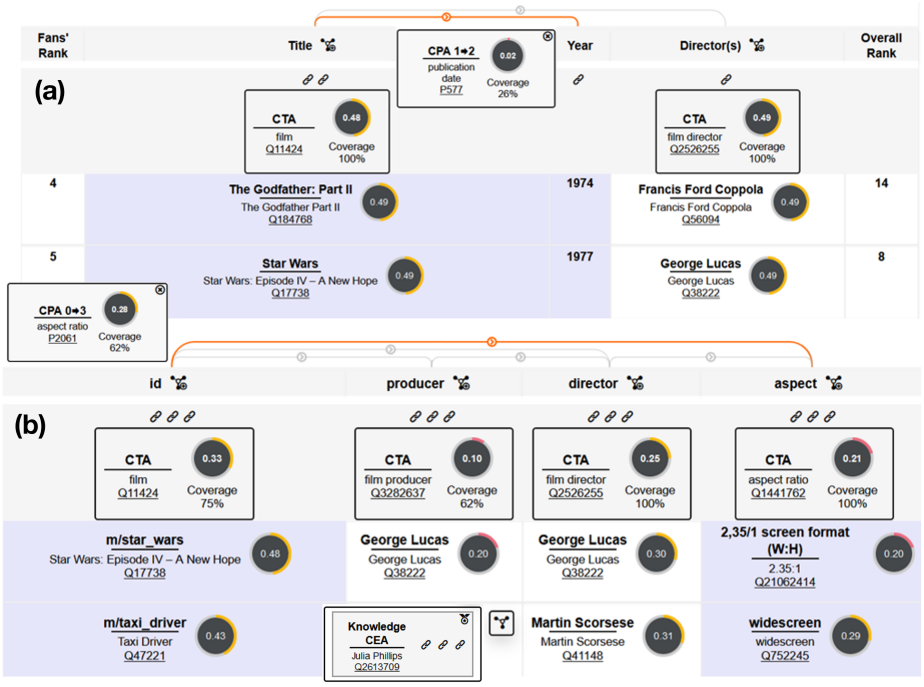

Fig. 1. DAGOBAH UI depicting the semantic annotations on the SemTab and Movie tables with the associated confidence scores.

table topology (orientation, header, etc.) which are crucial for the annotation process.

The user can then launch the semantic annotation process. The results of this process on Table *77694908_0_6083291340991074532.csv* (from SemTab 2019 dataset [4]) are presented in Figure 1.a. Three tasks are carried out. Cell-Entity Annotation (CEA) aims at associating each cell of the table with an entity of the KG. In DAGOBAH UI, the CEA results are presented together with the original mentions. For example, "Star Wars" has been annotated with the Wikidata entity Q177738 (Star Wars: Episode IV - A New Hope). Column-Type Annotation (CTA) aims to map each column with an entity type. In the user interface, these annotations are presented in the upper part of the table (in the headers). In the example, the system has annotated the "Title" column with the entity Q11424 (film). Columns-Property Annotation (CPA) seeks to associate pairs of table columns with a property of the KG. The relationships found are symbolised by links at the top of the table. When the user clicks on a link, the associated Wikidata property is displayed. In the example, the relationship P577 (publication date) has been identified between the columns "Title" and "Year". Figure 1.b shows the annotation results for another Web table about movies generated partly from Wikidata. This example illustrates the power of the semantic elevation enabled by the annotation process. Indeed,

the system found that "Star Wars" in the SemTab table and "m/star_wars" in the Movie table actually denote the same entity: Q177738 (Star Wars: Episode IV - A New Hope). This data reconciliation capability is particularly interesting for use cases involving heterogeneous datasets.

3.2 Table Enrichment with KGs

The life cycle of tabular data and the coverage of background knowledge represented in open or enterprise KGs can vary. This situation generates differences between tables published by organisations and open or enterprise KGs that are continuously curated: information is missing, dimensions are eluded since they are deemed useless by the data producer, etc. However, in many use cases (e.g. dataset search, profiling and recommendation), the completeness and richness of the data have positive effects and are desirable qualities. Once annotated, tables can be enriched with additional elements from the supporting KG. For example, missing values can be filled and new columns can be appended using the KG background knowledge. The enriched table can then be exported by the user for subsequent analysis. Figure 2 shows DAGOBAH UI modal window for selecting new columns to add to the Movie table. SPARQL queries are used to identify the most representative properties (e.g. cast member, cost, award received, etc.) of the entities in the column for which the user has requested suggestions (button next to the header). In the example, after selecting the P1476 (title) property, the user interface allows to preview the added values before confirming the operation.

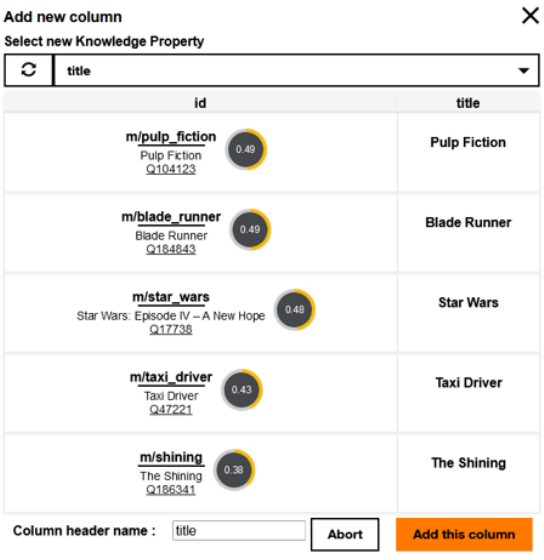

Fig. 2. Modal window for choosing new columns to add to the Movie table.

4 Demonstration

The demonstration at the conference will showcase the annotation and enrichment features of DAGOBAH UI. This will be done on different scenarios and datasets from various sectors such as creative industries, news and sports, and the pharmaceutical domain. Attendees will be able to discover how DAGOBAH UI can automatically interpret a table via CEA, CTA and CPA tasks. Table enrichment features will also be presented to show how the tool can generate richer and more complete tables. We provide a demonstration video on YouTube at `https://tinyurl.com/dagobah-ui-demo`

Our future work includes the development of new features around KG enrichment from tables. As discussed in this paper, tables can benefit from KGs through cell completion and table expansion with new columns. Conversely, tables are also a great source of dormant knowledge that can be leveraged to enrich open or enterprise KGs. To this aim, DAGOBAH UI will enable to export the annotations as RDF triples to refine KGs. Future work also includes the evaluation of DAGOBAH UI usability with real users. To this aim, the availability of DAGOBAH UI within the company will allow to collect interesting feedbacks to progress on the adoption of STI tools.

References

1. Chabot, Y., Monnin, P., Deuzé, F., Huynh, V., Labbé, T., Liu, J., Troncy, R.: A Framework for Automatically Interpreting Tabular Data at Orange. In: ISWC Posters, Demos and Industry Tracks. CEUR Workshop Proceedings, vol. 2980 (2021)
2. Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L., Kacprzak, E., Groth, P.: Dataset search: a survey. VLDB Journal **29**(1), 251–272 (2020)
3. Cremaschi, M., Rula, A., Siano, A., De Paoli, F.: Mantistable: A tool for creating semantic annotations on tabular data. In: European Semantic Web Conference (ESWC). pp. 18–23 (2019)
4. Hassanzadeh, O., Efthymiou, V., Chen, J., Jiménez-Ruiz, E., Srinivas, K.: SemTab 2019: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching Data Sets. Zenodo (2019), `https://doi.org/10.5281/zenodo.3518539`
5. Huynh, V.P., Liu, J., Chabot, Y., Deuzé, F., Labbé, T., Monnin, P., Troncy, R.: DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data. In: SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (2021)
6. Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K., Cutrona, V.: Results of semtab 2020. In: SemTab 2020: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching. vol. 2775, pp. 1–8 (2020)
7. Lehmberg, O., Ritze, D., Meusel, R., Bizer, C.: A Large Public Corpus of Web Tables containing Time and Context Metadata. In: 25^{th} International Conference Companion on World Wide Web (WWW Companion). pp. 75–76 (2016)
8. Nguyen, P., Yamada, I., Kertkeidkachorn, N., Ichise, R., Takeda, H.: SemTab 2021: Tabular Data Annotation with MTab Tool. In: SemTab 2021: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (2021)
9. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communication of the ACM **57**(10), 78–85 (2014)