DAGOBAH: Table and Graph Contexts for Efficient Semantic Annotation of Tabular Data

Viet-Phi Huynh[1], Jixiong Liu[1,2], Yoan Chabot[1], Frédéric Deuzé[1],
Thomas Labbé[1], Pierre Monnin[1], and Raphaël Troncy[2]

[1] Orange, France
yoan.chabot@orange.com
[2] EURECOM, Sophia Antipolis, France
raphael.troncy@eurecom.fr

Abstract. In this paper, we present the latest improvements of the DAGOBAH system that performs automatic pre-processing and semantic interpretation of tables. In particular, we report promising results obtained in the SemTab 2021 challenge thanks to optimisations in lookup mechanisms and new techniques for studying the context of nodes in the target knowledge graph. We also present the deployment of DAGOBAH algorithms within the Orange company via the TableAnnotation API and a front-end DAGOBAH user interface. These two access methods enable to accelerate the adoption of Semantic Table Interpretation solutions within the company to meet industrial needs.

Keywords: Semantic Table Interpretation · DAGOBAH · SemTab

1 Introduction

Tables constitute a major source of knowledge since large parts of companies internal repositories and Web pages are represented in tabular formats. Hence, there is a strong interest in Semantic Table Interpretation (STI), *i.e.*, methods that automatically interpret tables with the support of a knowledge graph (KG) by associating each cell with an individual (Cell-Entity Annotation, CEA), each column with a class (Column-Type Annotation, CTA), and each pair of columns with a property (Columns-Property Annotation, CPA) [3, 5, 8, 9]. Such annotations can then be used for various use cases ranging from dataset indexing and recommendation to knowledge graph refinement.

The DAGOBAH algorithms, conjointly developed by Orange and EURECOM, have been evaluated on all editions of the SemTab challenge and they have now reached a maturity level that allows to address industrial use cases related to STI and inherently present within Orange. Indeed, Orange is a multinational company operating in a wide range of business domains (*e.g.*, telecommunications, banking, multimedia content, cybersecurity). Consequently, Orange produces a high volume of heterogeneous tabular data. Thanks to STI techniques, these data could be strategically leveraged, *e.g.*, by structuring dormant knowledge, thus making it actionable and usable through Q/A engines [2].

Both the SemTab 2021 challenge and the aforementioned industrial needs motivated the following improvements in the entity scoring algorithm of the DAGOBAH SL 2020 system [7]:

- An enhancement of indexing and entity matching strategies to improve the lookup quality as well as the lookup coverage;
- A better representation and disambiguation of entities in exploiting more efficiently their contexts in the KG;
- An improved and flexible entity scoring that leverages both local information and global table information.

These specific improvements yield the DAGOBAH SL 2021 system which is detailed in Section 2. We report on the results of our experimental evaluation as well as the insights we gain from the challenge leaderboard in Section 3. Section 4 introduces our efforts towards the usability of DAGOBAH within Orange via the TableAnnotation RESTful API and a new front-end DAGOBAH user interface. Finally, we reflect on the SemTab challenge and the adoption of STI solutions within the enterprise in Section 5.

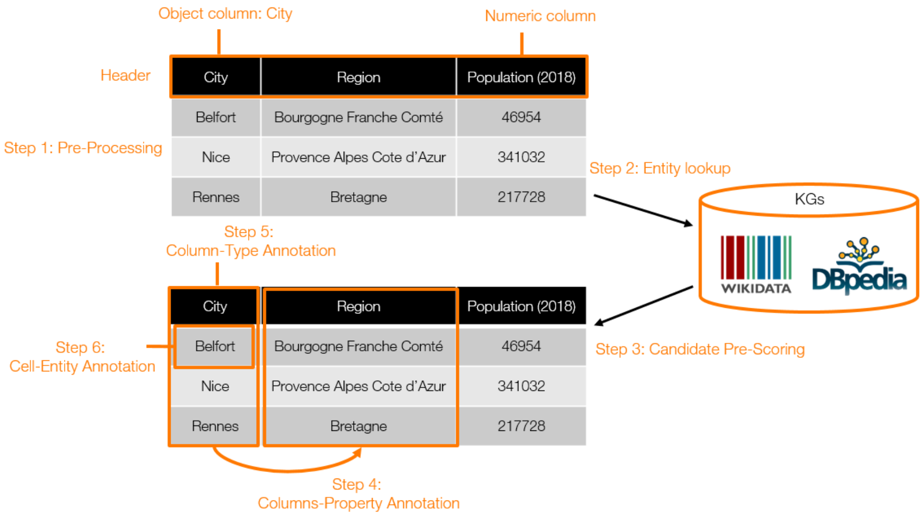
2 DAGOBAH SL 2021: Optimised Lookup, Contexts in Knowledge Graphs and Flexible Entity Scoring

DAGOBAH provides an end-to-end process that annotates relational tables with constituents of a KG such as Wikidata. Its workflow consists of the four sequential steps depicted in Figure 1. Given a relational table as input, the preprocessing determines table metadata and annotation targets (Section 2.1). The entity lookup module then collects candidates from the KG for each target cell of the table (Section 2.2). The pre-scoring module evaluates each candidate to determine a confidence score (Section 2.3). Next, Columns-Property Annotation (CPA) and Column-Type Annotation (CTA) are carried out (Section 2.4). Finally, Cell-Entity Annotation (CEA) is performed in order to compute the final entity scores taking CTA and CPA into account (Section 2.4).

2.1 Table Pre-Processing

In real use cases, annotating tables is complex because of little prior knowledge about their structure and content. Therefore, pre-processing tables can facilitate later annotations. In this view, DAGOBAH pre-processing steps generates metadata about a table via four main tasks: orientation detection, header detection, key column detection[3] and column primitive typing. This primitive typing detects named entities (*e.g.*, Location, Organization, Person), literals with

[3] Only single key column is currently supported.



City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

Fig. 1. Overview of the DAGOBAH annotation workflow.

units (*e.g.*, Distance, Speed, Temperature), or miscellaneous literals (*e.g.*, Email, URL, IP Address). The pre-processing step was particularly helpful for the BioDivTab [1] and GitTables [6] datasets (Section 3.2).

2.2 Entity Lookup

The pre-processing step helps to identify table's columns eligible for entity lookup based on their primitive types[4]. Given a cell e_m in such a column, the entity lookup step retrieves a set of relevant candidate entities $\mathcal{E}_c(e_m)$ from a target KG. The lookup service of DAGOBAH is based on Elasticsearch and currently supports two KGs for which indexes have been built: Wikidata and DBpedia.

Wikidata entities. The lookup service collects items and properties with their labels and aliases in all languages. To increase lookup coverage, aliases of each entity are enriched with the values of 11 additional properties such as P2561 (name), P1705 (native label), or P742 (pseudonym).

DBpedia entities. The lookup service collects English resources with their labels in all languages. To increase lookup coverage, labels are enriched with the values of 25 alias properties such as abbreviation, birthName, or originalTitle. In addition, labels and aliases of all redirected entities are also included.

We average the character-based and token-based edit distances[5] to evaluate the similarity between a cell mention and the set of labels of each candidate

[4] Since target cells are given in the SemTab challenge, this feature is not used in our experiments.

[5] https://github.com/seatgeek/thefuzz

entity. This helps to solve the "out of order problem" where a string can have different orders for its substrings (*e.g.*, "Elon Musk" and "Musk Elon").

2.3 Candidate Pre-Scoring

The pre-scoring step evaluates with a preliminary score the relevance of a candidate entity $e_c \in \mathcal{E}_c(e_m)$ for a table cell e_m:

$$PSc(e_c, e_m) = Sc_{context}(\mathcal{N}_{graph}(e_c), \mathcal{N}_{table}(e_m)) \times e^{\gamma(Sc_{sim}(e_c,e_m)-1)} \quad (1)$$

This pre-score is the product of a context factor and a literal factor $Sc_{sim}(e_c, e_m)$. The latter returns the highest Levenshtein-based matching ratio between the cell and the label and aliases of the candidate. Aliases are penalized with a ratio weighted by 0.9 since we consider labels to be more important. The amplification factor $\gamma \in \mathbb{N}^+$ determines the importance of the textual similarity. We empirically observed that 2 was an appropriate amplification factor for the SemTab challenge.

The improvements in DAGOBAH SL 2021 mainly concern the context factor defined as follows:

$$Sc_{context}(\mathcal{N}_{graph}(e_c), \mathcal{N}_{table}(e_m)) = \frac{\sum_i w_i \times sn_i}{\sum_i w_i} \quad (2)$$

where $\mathcal{N}_{table}(e_m)$ is the set of neighboring cells in the same row as e_m and $\mathcal{N}_{graph}(e_c)$ is the set of neighboring nodes of e_c in the KG[6]. For each neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$, sn_i is its neighborhood matching score w.r.t. $\mathcal{N}_{graph}(e_c)$.

DAGOBAH SL 2021 solves two issues related to the context score calculation with respect to DAGOBAH SL 2020:

Expensive evaluation. Each sn_i was evaluated by iterating over all context nodes in $\mathcal{N}_{graph}(e_c)$ to find the best matching node. Hence, a performance bottleneck arose when scoring a generic entity with millions of edges in the KG. For example, let's consider the cell "Belfort" in Figure 1 and the Wikidata candidate entity Q171545. To check whether the neighboring cell "Bourgogne Franche Comté" is in the context of Q171545, we performed a comparison with each of the $\sim$ 1000 nodes in $\mathcal{N}_{graph}(Q171545)$, including Q142 (France), Q3371185 (Paul Faivre), etc. (Figure 2a).

One-hop graph contexts. $\mathcal{N}_{graph}(e_c)$ consisted of nodes only one hop away from e_c. Consequently, a neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ matching with a node two hops away from e_c was not considered in the context of e_c. For example, given the one-hop context of Q171545 (Belfort) in Figure 2a, we wrongly considered that Bourgogne Franche Comté had no relation with Belfort whereas it is the region of Territoire de Belfort (French department) whose capital is Belfort.

DAGOBAH SL 2021 improves both the efficiency and the expressiveness of the context score by avoiding the exhaustive scoring and exploiting more expressive contexts for an entity via two-hop predicate paths.

[6] Neighboring nodes are connected to e_c via predicate paths in the KG, regardless of the predicate direction.

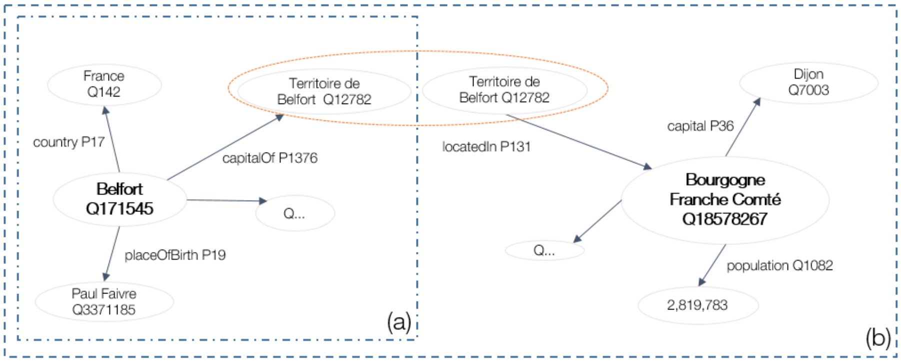


Fig. 2. Graph context of entity Q171545 (Belfort) in Wikidata. (a) One-hop graph context of Q171545. (b) Graph context is expanded by sub-graph intersection.

Exploiting the Context of Knowledge Graph Entities. The neighborhood matching score sn_i in Equation (2) indicates whether a neighboring cell n_i of e_m matches with a neighboring node of e_c. Loosely speaking, computing sn_i comes down to searching a candidate entity for n_i in $\mathcal{N}_{graph}(e_c)$ and to assessing its similarity. In our running example, Q18578267 is a candidate for cell "Bourgogne Franche Comté" in the two-hop context $\mathcal{N}_{graph}(Q171545)$ (Figure 2b). From this observation, we propose the following way to efficiently compute sn_i. The entity lookup step (Section 2.2) outputs candidate entities $\mathcal{E}_c(e_m)$ for a target cell e_m but also candidate entities $\mathcal{E}_c(n_i)$ for its neighboring cells n_i. Hence, we check if a candidate entity $e_i \in \mathcal{E}_c(n_i)$ is in $\mathcal{N}_{graph}(e_c)$. In that case, sn_i is simply calculated by comparing the labels of the neighboring cell n_i and the matching node e_i, which avoids additional comparisons with other nodes in $\mathcal{N}_{graph}(e_c)$.

To check if $e_i \in \mathcal{E}_c(n_i)$ is in $\mathcal{N}_{graph}(e_c)$, we actually check if e_i is connected to e_c by a predicate path in the KG. We chose to compute such predicate paths since they are useful in the soft context scoring. To efficiently find predicate paths between e_c and e_i, we extract the one-hop subgraphs $\mathcal{G}_{e_c}$ and $\mathcal{G}_{e_i}$ around e_c and e_i. If an intermediate node v is present in both $\mathcal{G}_{e_c}$ and $\mathcal{G}_{e_i}$, the paths pointing to v in the two sub-graphs are concatenated. In our running example, we find the following predicate path: $\mathsf{Belfort} \xrightarrow{\mathsf{capitalOf}} \mathsf{Territoire\ de\ Belfort} \xrightarrow{\mathsf{locatedIn}}$ $\mathsf{Bourgogne\ Franche\ Comté}$. Since we only consider one-hop subgraphs, paths can have a maximum length of two hops. This approach allows to enrich the information about an entity by including not only direct neighbors but also indirect neighbors two hops away. Such enhanced graph contexts increase the chance for a neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ to match, and thus make the context score more precise. We argue that, in the context of STI with Wikidata, paths longer than 2 hops are often noisy and meaningless, and thus can have a negative impact on the context score.

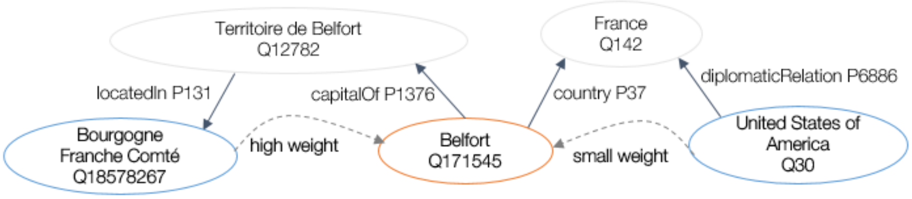


Fig. 3. Neighboring nodes of Belfort (Q171545) contribute differently to its information content.

Soft Context Scoring. In Equation (2), neighborhood matching scores sn_i are weighted to compute the ultimate score of an entity. Indeed, each neighboring cell $n_i \in \mathcal{N}_{table}(e_m)$ contributes differently to the annotation of the target cell e_m with a weight w_i defined in Equation (3):

$$w_i = \frac{\overbrace{se_i}^{(3a)}}{\underbrace{\sqrt{d(col_i)+1}}_{(3b)}} \times \overbrace{cnt(col_i)}^{(3c)} \times \overbrace{\tau(e_i)}^{(3d)} . \tag{3}$$

(3a) Cells containing entities should be more important than cells containing literals (*e.g.*, date, measurement with/without unit, number) since there is a lack of literal disambiguation methods (*e.g.*, date-time normalization, unit detection/normalization/conversion). That is why, we set se_i to 1.0 if the neighboring cell n_i contains an entity, and to 0.15 if n_i contains a literal.

(3b) A neighboring cell on the left side of the table has more chance to be a meaningful context for the target cell. Hence, $d(col_i)$ is the distance between column col_i and the first object column of the table.

(3c) Cells n_i from a neighboring column highly connected to the target column should have a greater weight in the context. Hence, we take into account the connectivity $cnt(col_i)$ of the neighboring column w.r.t. the target column, defined as the highest occurrence of a relation potentially found between the two columns.

(3d) Neighboring nodes of the candidate entity e_c in $\mathcal{N}_{graph}(e_c)$ may provide different information content as some neighbors can be "semantically closer" to e_c than others. To illustrate, consider the 2-hop context of Q171545 (Belfort) depicted in Figure 3. Q18578267 (Bourgogne Franche Comté) is more relevant than Q30 (United States of America) since the path Belfort $\xrightarrow{\text{capitalOf}}$ Territoire de Belfort $\xrightarrow{\text{locatedIn}}$ Bourgogne Franche Comté is more informative than Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America. To quantify this, we rely on the so-called truth value $\tau(e_i)$ [4] of a neighboring node e_i, which can be seen as the discriminative capacity of the associated path

$\tau(e_c \xrightarrow{p_1} v \xrightarrow{p_2} e_i)$, and is defined as follows:

$$\tau(e_i) = \tau(e_c \xrightarrow{p_1} v \xrightarrow{p_2} e_i) = \frac{1}{1 + log(g(v))} \tag{4}$$

where $g(v)$ is the generality of the intermediate node v, *i.e.*, the number of its incoming and outcoming edges in the KG. Note that direct neighbors (or 1-hop paths) always get the highest truth value 1.0.

2.4 Annotation Tasks

Columns-Property Annotation. The CPA task outputs the most suitable semantic relation r for an ordered pair of columns. We adopt a majority voting strategy that relies on the occurrence and accumulated confidence score over rows for r. The interested reader can refer to [7] for additional details. Note that, accordingly to Section 2.3, r can be one-hop (*i.e.*, $\xrightarrow{p}$), unidirectional 2-hop (*i.e.*, $\xrightarrow{p1}\xrightarrow{p2}$ or $\xleftarrow{p1}\xleftarrow{p2}$), or bidirectional 2-hop (*i.e.*, $\xrightarrow{p1}\xleftarrow{p2}$ or $\xleftarrow{p1}\xrightarrow{p2}$).

Column-Type Annotation. The CTA task aims to identify the most representative and specific type for a target column. Types of candidate entities in this column are collected and a majority voting strategy is used to determine the most precise type (see [7] for more details on type enrichment and score calculation).

Cell-Entity Annotation. The CEA task selects for a table cell e_m the most relevant entity among candidate entities $e_c \in \mathcal{E}_c(e_m)$ retrieved from the KG. This step leverages both entity pre-scoring and information given by CTA and CPA to compute the final score of candidate entities. Indeed, the pre-scoring of a candidate entity e_c only considers its local information, *i.e.*, the row it belongs to. Column type output by CTA and column pair relations output by CPA allow to consider table's global information. Hence, the final score $Sc(e_c, e_m)$ of a candidate entity e_c is computed as follows:

$$Sc(e_c, e_m) = \frac{(PSc(e_c, e_m) + \alpha \times score_{CTA} + \beta \times \overline{score_{CPA}})}{1 + \alpha + \beta} \tag{5}$$

If e_c belongs to the type output by CTA for its column, then $score_{CTA}$ is equal to the score of this type, otherwise it is set to 0. In $\overline{score_{CPA}}$, we average the scores of the relations output by CPA for column pairs involving the column of e_c. For each relation, if e_c belongs to its domain or its range (depending on the relation orientation), then we consider the score of this relation, else it is set to 0. To strengthen (resp. weaken) a frequent (resp. unusual) CTA/CPA in the update of $Sc(e_c, e_m)$, a coefficient α (resp. β) is employed and defined as $\frac{occurrence(CTA)}{2}$ (resp. $\frac{\overline{occurrence(CPA)}}{2}$). Note that the occurrence of CTA/CPA is divided by 2 to prioritize the pre-scoring $PSc(e_c, e_m)$.

3 Experiments

3.1 Settings

To evaluate one-hop and two-hop graph contexts as well as the soft context scoring described in Section 2, we consider the four experimental settings:

Setting 1 The context score of an entity is computed using only its one-hop neighboring graph. Weights w_i do not follow Equation (3) but are set to 1.0 for entities and 0.15 for literals.

Setting 2 The context score of an entity is computed using its two-hop neighboring graph. Weights w_i do not follow Equation (3) but are set to 1.0 for 1-hop neighbors, 0.25 for 2-hop neighbors, and 0.15 for literals.

Setting 3 The context score of an entity is computed using its two-hop neighboring graph. Weights w_i follow Equation (3). This setting allows to assess if richer contexts and stricter scoring lead to better annotation.

Setting 4 This setting restricts Setting 3 to 1-hop and unidirectional 2-hop predicate paths in graph contexts. This allows to evaluate the impact of bidirectional paths which are often less informative or noisy but may be helpful in some cases.

3.2 Results

Experimental evaluation. We provide an experimental evaluation of the four aforementioned settings in Table 1. It should be noted that DAGOBAH is continuously improved. Hence, results of this evaluation are based on the current version of DAGOBAH but we also report results submitted to the SemTab challenge in gray cells for comparison. To validate the modifications proposed in Sections 2.2 and 2.3, we include the scores of the DAGOBAH 2020 system on tables annotated with Wikidata in Round 1. Submission Settings {1,2,3,4}* are similar to Settings {1,2,3,4} but slightly differ in scores and weights initialization. This does not change CEA scores but impacts CTA performances. Indeed, CTA is highly sensitive to entity scores and taxonomy weights to select the most fine-grained type among the many possible (direct and ancestor) types of entities.

DAGOBAH achieves a high performance on synthetic datasets (Round 2) whereas high-quality manually-curated datasets with complex table patterns are more difficult to annotate (Rounds 1 and 3). In HardTable, no gain is brought by using a richer graph context or a more flexible scoring. This can be explained since tables are almost fully represented in the target KG and columns can be disambiguated from their contents. On the contrary, BioTable provides remarkable ambiguities with content overlaps between columns that hinder their disambiguation (*e.g.*, column “Gene” can be mistaken with column “Protein”). Hence, annotation seems to benefit from richer graph contexts. In BioDivTable, Setting 4 obtains the lowest scores whereas Setting 1 is comparable to Setting 3. We suppose that unidirectional 2-hop predicate paths may be noisy or not correctly considered, leading to the lowest score of Setting 4.

In general, Settings 2, 3, and 4 are more precise for CEA than Setting 1. Hence, 2-hop graph contexts bring useful information. The better performance of Settings 3 and 4 compared with Setting 2 shows the effectiveness of soft context scoring. We notice that Setting 3 achieves similar performances to Setting 4 which can be interpreted as follows. First, unidirectional paths (*i.e.*, $\xrightarrow{p_1}\xrightarrow{p_2}$ and $\xleftarrow{p_1}\xleftarrow{p_2}$) bring enough information and allow to obtain equal results compared with considering both unidirectional and bidirectional paths. Second, the influence of noisy bidirectional paths (*e.g.*, Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America) is limited by the soft context scoring which prevents a degradation in the annotation quality. This allows useful bidirectional paths to contribute positively in the entity score. It can be observed that CTA and CPA performances are not as high as expected in most datasets despite CEA good performances. The development of better strategies for type and relation selection will be the subject of future works.

Table 1. Comparison of experimental settings and performance of the DAGOBAH system in Rounds 1, 2, and 3 of the SemTab 2021 challenge (in gray cells). "F1" stands for F1-score, "P" stands for Precision. Best results between settings are in bold.

Dataset	System Setting	CTA		CEA		CPA	
		F1	P	F1	P	F1	P
Round 1 – WDTable	Setting 1	**0.793**	**0.793**	0.913	0.913	-	-
	Setting 2	0.790	0.790	0.923	0.923	-	-
	Setting 3	0.783	0.783	**0.926**	**0.926**	-	-
	Setting 4	0.783	0.783	0.924	0.924	-	-
	DAGOBAH 2020	0.743	0.743	0.830	0.841	-	-
	Setting 2*	0.832	0.832	0.923	0.923	-	-
Round 1 – DBPTable	Setting 1	0.25	0.25	0.935	0.935	-	-
	Setting 2	0.27	0.27	0.946	0.946	-	-
	Setting 3	**0.274**	**0.274**	**0.947**	**0.947**	-	-
	Setting 4	**0.274**	**0.274**	**0.947**	**0.947**	-	-
	Setting 2*	0.422	0.424	0.945	0.946	-	-
Round 2 – BioTable	Setting 1	0.874	0.874	0.882	0.882	0.898	**0.901**
	Setting 2	0.911	0.911	0.916	0.916	**0.899**	0.899
	Setting 3	0.915	0.915	0.950	0.951	**0.899**	0.899
	Setting 4	**0.916**	**0.916**	**0.970**	**0.970**	**0.899**	0.899
	Setting 4*	0.916	0.916	0.970	0.970	0.899	0.899
Round 2 – HardTable	Setting 1	**0.968**	**0.969**	0.975	**0.976**	**0.996**	**0.997**
	Setting 2	**0.968**	**0.969**	**0.976**	**0.976**	**0.996**	**0.997**
	Setting 3	**0.968**	**0.969**	**0.976**	**0.976**	**0.996**	**0.997**
	Setting 4	**0.968**	0.968	**0.976**	**0.976**	**0.996**	**0.997**
	Setting 3*	0.976	0.976	0.975	0.976	0.996	0.996
Round 3 – BioDivTable	Setting 1	0.338	0.339	0.619	0.64	-	-
	Setting 2	0.335	0.335	0.60	0.62	-	-
	Setting 3	**0.344**	**0.345**	**0.62**	**0.641**	-	-
	Setting 4	0.343	0.343	0.475	0.491	-	-
	Setting 4*	0.381	0.382	0.496	0.497	-	-
Round 3 – HardTable	Setting 3*	0.99	0.99	0.974	0.974	0.991	0.995
Round 3 – GitTables DBP	Pre-processing + Mapping	0.07	0.117	-	-		-
Round 3 – GitTables SCH	Pre-processing + Mapping	0.183	0.185	-	-	-	-

BioDivTab and GitTables datasets. Note that for BioDivTab and GitTables datasets, we adapted the DAGOBAH algorithms presented in this paper. Indeed, for BioDivTab, primitive types output by the pre-processing step were used to discriminate object and literal columns. A column contains literal values if its primitive type is numerical, date-time, unit, or miscellaneous. Otherwise, it is considered as an object column containing entity mentions that are passed to the lookup module. For GitTables, primitives types are manually mapped to Schema.org and DBpedia Ontology classes.

4 Interpreting Tabular Data at Orange

To quickly improve the relevance of DAGOBAH w.r.t. real use cases, we adopted a Test & Learn approach. In this view, DAGOBAH algorithms have been made available to other collaborators inside the company since the early stages of the project via the TableAnnotation API and the DAGOBAH UI.

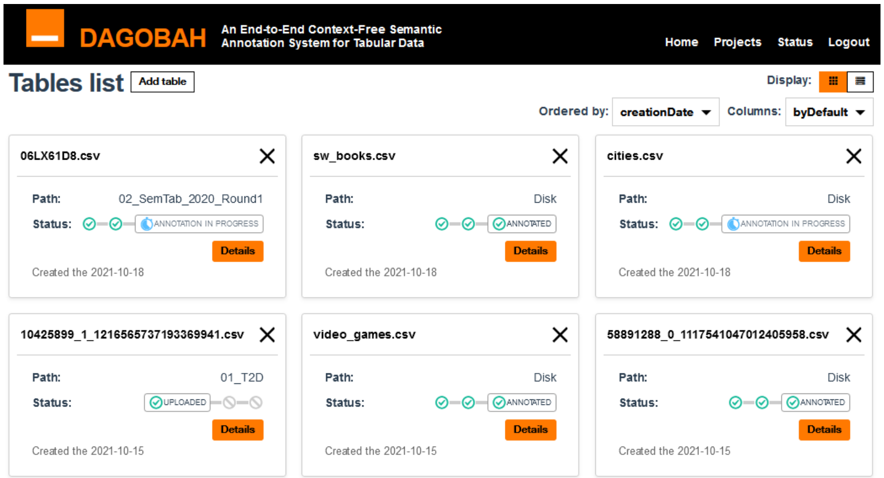
TableAnnotation API. This RESTful API is deployed on the Orange Developer portal[7]. It provides pre-processing and annotation services for tables, as well as lookup services to disambiguate mentions and retrieve corresponding Wikidata or DBpedia entities. This API is accessible to all company's R&D teams and business units upon request. We plan on extending the API access to external users in the future.

DAGOBAH UI. This interface allows non-developers and non-AI experts to use the TableAnnotation API resources on their tables and to visualise the results in an intelligible and ergonomic form. This interface enables users to load new tables into their annotation project (Figure 4a), start the pre-processing and semantic annotation tools, and visualise the results (Figure 4b). As DAGOBAH UI is a powerful tool to demonstrate the value of STI techniques within the company or with external prospects, future developments will focus on features such as:

- KG enrichment with table elements not present in the KGs;
- Table enrichment with KGs by filling in missing values or adding new columns;
- Interactive visualisation of the target KG, annotations resulting from CEA, CTA, and CPA, and new triples that can be generated from the table.

This front-end user interface allows collaborators to understand the interest of STI solutions. Conversely, the DAGOBAH project team can identify challenges associated with their needs, which is a valuable input for the project roadmap. Although deployment and adoption of STI methods in Orange are still in their infancy, tests on different use cases have been taking place for over a year via the TableAnnotation API, which has received more than 200,000 API calls. Several areas are being prioritised including entertainment (*e.g.*, VOD catalog), data governance, and health.

[7] https://developer.orange.com



Tables list

Ordered by: creationDate Columns: byDefault

06LX61D8.csv
Path: 02_SemTab_2020_Round1
Status: ANNOTATION IN PROGRESS
Created the 2021-10-18

sw_books.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-18

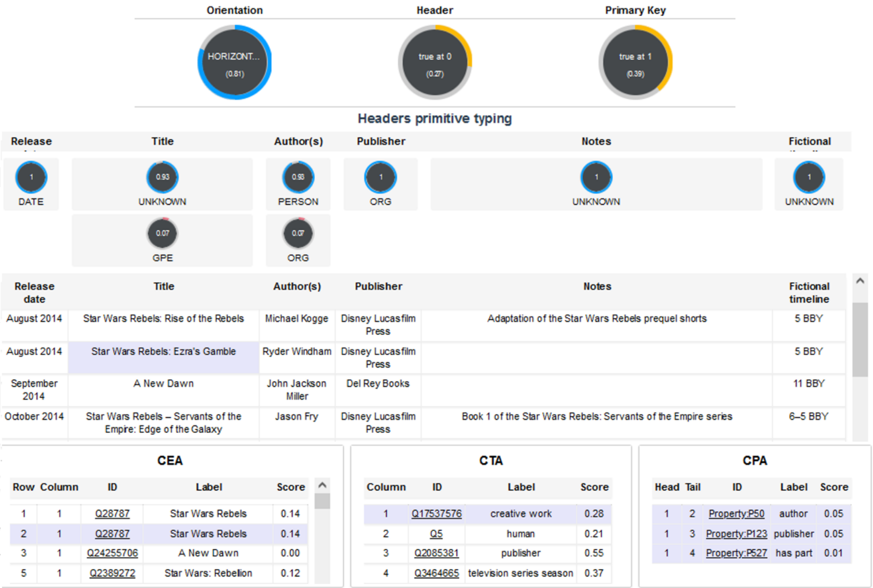
cities.csv
Path: Disk
Status: ANNOTATION IN PROGRESS
Created the 2021-10-18

10425899_1_1216565737193369941.csv
Path: 01_T2D
Status: UPLOADED
Created the 2021-10-15

video_games.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-15

58891288_0_1117541047012405958.csv
Path: Disk
Status: ANNOTATED
Created the 2021-10-15

(a) DAGOBAH UI allows to define projects and load tables from the local file system or pre-loaded tables from gold standards (*e.g.*, T2D, SemTab).



Release date	Title	Author(s)	Publisher	Notes	Fictional timeline
August 2014	Star Wars Rebels: Rise of the Rebels	Michael Kogge	Disney Lucasfilm Press	Adaptation of the Star Wars Rebels prequel shorts	5 BBY
August 2014	Star Wars Rebels: Ezra's Gamble	Ryder Windham	Disney Lucasfilm Press		5 BBY
September 2014	A New Dawn	John Jackson Miller	Del Rey Books		11 BBY
October 2014	Star Wars Rebels – Servants of the Empire: Edge of the Galaxy	Jason Fry	Disney Lucasfilm Press	Book 1 of the Star Wars Rebels: Servants of the Empire series	6–5 BBY

CEA

Row	Column	ID	Label	Score
1	1	Q28787	Star Wars Rebels	0.14
2	1	Q28787	Star Wars Rebels	0.14
3	1	Q24255706	A New Dawn	0.00
5	1	Q2389272	Star Wars: Rebellion	0.12

CTA

Column	ID	Label	Score
1	Q17537576	creative work	0.28
2	Q5	human	0.21
3	Q2085381	publisher	0.55
4	Q3464665	television series season	0.37

CPA

Head	Tail	ID	Label	Score
1	2	Property:P50	author	0.05
1	3	Property:P123	publisher	0.05
1	4	Property:P527	has part	0.01

(b) DAGOBAH UI allows to access pre-processing and annotation details for a table. At the top, information generated by the pre-processing (*e.g.*, orientation, header) and the cleaned table are displayed. An interactive view shows CEA, CTA and CPA annotations at the bottom.

Fig. 4. Features of DAGOBAH UI.

5 Discussion

The datasets provided for the SemTab 2021 challenge have extended the spectrum of difficulties for STI methods by including new target domains (*e.g.*, biomedical and Git data), combining KGs, or requiring schema-only annotations (Schema.org and DBpedia Ontology). These difficulties have enabled us to enhance the DAGOBAH system annotation strategies, *e.g.*, leveraging preprocessing primitive types, and using enriched graph contexts. Nonetheless, new research directions can still be explored to embrace the heterogeneity of table types that are published on the Web:

- Table structure and inner-relationships: table orientation, nested cells, layout concatenation, multi-valued cells, subjects split into several columns (*e.g.*, first and last names of a person), etc.
- Out-of-KG data: entities not present in a given KG, which is often the case with companies specific data.

It is noteworthy that out-of-KG data started to be addressed in Round 3 with the GitTables dataset that required schema-only annotations from Schema.org and the DBpedia Ontology. However, this task was not fully consistent with the CTA definition adopted by the community since ground truth annotations mixed types (classes) and properties. These heterogeneous annotations may lead to inconsistent evaluations. To address the aforementioned dimensions, the DAGOBAH team is working on hard datasets that will bring new challenges to the community.

6 Conclusion

In this paper, we have presented the improvements made on the DAGOBAH system. With optimised lookups, richer graph contexts, and soft scoring, DAGOBAH obtained high performances during the SemTab 2021 challenge. Our future work aims at increasing the annotation accuracy for tables with non-explicit or highly ambiguous mentions. We ambition to leverage dictionaries providing abbreviations or acronyms. To ensure the genericity of our approach, such dictionaries should be built from huge amounts of documents and be applicable on various datasets. When a majority of unmatchable mentions are present in a column, embedding-based approaches could complement context scoring strategies. In this view, we believe that approaches based on language models can become a good asset in the most challenging cases.

Acknowledgments. The authors would like to thank Christophe Sarthou-Camy and Guillaume Jourdain for their invaluable contribution to the development of DAGOBAH UI.

References

1. Abdelmageed, N., Schindler, S., König-Ries, B.: BiodivTab: A Tabular Benchmark based on Biodiversity Research Data. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab) (2021)
2. Chabot, Y., Monnin, P., Deuzé, F., Huynh, V., Labbé, T., Liu, J., Troncy, R.: A Framework for Automatically Interpreting Tabular Data at Orange. In: 20^{th} International Semantic Web Conference (ISWC), Posters, Demos and Industry Tracks. CEUR Workshop Proceedings, vol. 2980 (2021)
3. Chen, S., Karaoglu, A., Negreanu, C., Ma, T., Yao, J., Williams, J., Gordon, A., Lin, C.: LinkingPark: An Integrated Approach for Semantic Table Interpretation. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2775, pp. 65–74 (2020)
4. Ciampaglia, G.L., Shiralkar, P., Rocha, L.M., Bollen, J., Menczer, F., Flammini, A.: Computational fact checking from knowledge networks. PloS one **10**(6) (2015)
5. Cremaschi, M., Avogadro, R., Barazzetti, A., Chieregato, D.: MantisTable SE: an Efficient Approach for the Semantic Table Interpretation. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2775, pp. 75–85 (2020)
6. Hulsebos, M., Demiralp, a., Groth, P.: Gittables: A large-scale corpus of relational tables. arXiv preprint arXiv:2106.07258 (2021), https://arxiv.org/abs/2106.07258
7. Huynh, V., Liu, J., Chabot, Y., Labbé, T., Monnin, P., Troncy, R.: DAGOBAH: enhanced scoring algorithms for scalable annotations of tabular data. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2775, pp. 27–39 (2020)
8. Nguyen, P., Yamada, I., Kertkeidkachorn, N., Ichise, R., Takeda, H.: Mtab4wikidata at semtab 2020: Tabular data annotation with wikidata. In: International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2775, pp. 86–95 (2020)
9. Shigapov, R., Zumstein, P., Kamlah, J., Oberländer, L., Mechnich, J., Schumm, I.: bbw: Matching CSV to Wikidata via Meta-lookup. In: Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K., Cutrona, V. (eds.) International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEUR Workshop Proceedings, vol. 2775, pp. 17–26 (2020)