# DAGOBAH: Annotation sémantique de données tabulaires par comparaison du contexte des tables et d'un graphe de connaissances

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

[1] Orange, France
[2] EURECOM, Sophia Antipolis, France

## Résumé

*Cet article présente les améliorations apportées à DAGOBAH, un système effectuant un pré-traitement automatique et une interprétation sémantique de données tabulaires en fonction d'un graphe de connaissances. Nous détaillons les optimisations des mécanismes de recherche de candidats et les nouvelles techniques d'étude du contexte des nœuds du graphe de connaissances cible qui nous ont permis d'obtenir les meilleures performances lors du challenge SemTab 2021 en terme de précision. Nous décrivons également le déploiement des algorithmes DAGOBAH au sein de l'entreprise Orange via l'API TableAnnotation et une interface utilisateur. Ces deux méthodes d'accès permettent d'accélérer l'adoption de solutions d'interprétation de tables au sein de l'entreprise pour répondre à des besoins industriels.*

## Mots-clés

*Interpretation sémantique de tables, DAGOBAH, SemTab*

## 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

Les données tabulaires constituent une source importante de connaissances, une grande partie des gisements internes des entreprises et du Web étant représentée sous cette forme. Par conséquent, il existe un vif interêt pour le domaine de l'interprétation automatique de données tabulaires (en anglais *Semantic Table Interpretation* ou STI). Ce domaine est caractérisé par le développement de méthodes d'interprétation automatique de tables à l'aide d'un graphe de connaissances via trois tâches principales. La tâche *Cell-Entity Annotation* (CEA) consiste à associer chaque cellule de la table avec une entité du graphe de connaissances. Par exemple, la mention "Belfort" de la Figure 1 sera annotée avec l'entité Q171545 (Belfort) du graphe de connaissances Wikidata. La tâche *Column-Type Annotation* (CTA) vise à annoter chaque colonne avec une classe. Par exemple, la première colonne "City" de la Figure 1 sera annotée avec l'entité Q484170 (commune française). Enfin, la tâche *Columns-Property Annotation* (CPA) vise à associer chaque paire de colonnes à une propriété. Par exemple, la relation entre les colonnes "City" et "Region" dans la Figure 1 serait la propriété P361 (fait partie de). Les annotations ainsi générées peuvent être utilisées dans de nombreux cas d'utilisation, de l'indexation des jeux de données et leur recommandation jusqu'à l'enrichissement de graphes de connaissances.

Les algorithmes de STI DAGOBAH, développés conjointement par Orange et EURECOM, ont été évalués lors des différentes éditions du challenge international SemTab [1] [8, 13, 14], colocalisé avec la conférence ISWC. Cet évènement centré sur les problématiques d'annotations de données tabulaires a attiré près de 50 équipes participantes au cours des trois dernières éditions. Comme démontré lors de ce challenge, nos outils ont atteint un niveau de maturité permettant de répondre à des problématiques industrielles dans le groupe. En effet, Orange est une multinationale opérant dans un grand nombre de domaines métiers (e.g. télécommunications, contenu multimédia, cybersécurité, etc.). Par conséquent, Orange produit de grands volumes de données tabulaires hétérogènes. A l'aide des techniques de STI, ces données peuvent être exploitées stratégiquement, par exemple, en structurant les connaissances dormantes dans ces données et en les rendant exploitables par le biais de moteurs de type questions-réponses [4].

Le challenge SemTab2021 et les besoins industriels mentionnés ci-dessus ont motivé des travaux de recherche qui constituent le cœur des algorithmes utilisés par le système DAGOBAH SL présenté en 2020 [11] et amélioré en

1. Semantic Web Challenge on Tabular Data to Knowledge Graph Matching : http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

2021 [10]. En particulier, le calcul de scores et le classement des entités ont été optimisés grâce à :

— Une amélioration des stratégies d'indexation et de mise en correspondance des entités pour améliorer la qualité et la couverture de la recherche de candidats (i.e. *lookup*) ;
— Une meilleure représentation et désambiguïsation des entités en exploitant plus efficacement leurs contextes (i.e. voisinage) dans le graphe de connaissances ;
— Un algorithme de notation (i.e. *scoring*) des entités amélioré et plus flexible exploitant à la fois des informations locales et des informations globales à la table étudiée.

Ces nouvelles contributions ont donné naissance au système DAGOBAH SL 2021 décrit dans la Section 3. Nous présentons les résultats de l'évaluation menée dans le cadre du challenge dans la Section 4. La Section 5 introduit les efforts autour de l'utilisabilité des systèmes d'annotation avec en particulier la mise à disposition d'une API REST `TableAnnotation` ainsi qu'une interface utilisateur nommée DAGOBAH UI. Enfin, la Section 6 offre des éléments de réflexions autour du challenge SemTab et de l'adoption des outils de STI au sein des entreprises.

# 2 Etat de l'art

L'approche courante pour réaliser la tâche de CEA consiste à effectuer des opérations de recherche syntaxique (e.g. Levenshtein), d'alignement d'ontologies ou d'exploitation de plongements [15]. La désambiguïsation des entités candidates est ensuite traitée comme une tâche de classement des candidats, en utilisant des heuristiques, des algorithmes tels que PageRank [9] ou des modèles basés sur les graphes [12]. Les principales approches sur le typage de colonnes (CTA) infèrent des classes à partir des entités produites par la tâche CEA. Des heuristiques plus ou moins complexes construites autour du vote majoritaire sont utilisées [17]. Enfin, l'extraction de relations (CPA) est généralement réalisée par la recherche de paires d'éléments en colonnes, *i.e.* types et entités préalablement choisis [19].

Récemment, le challenge SemTab a permis d'accélérer le développement des approches de STI. Une majorité d'entre elles prennent la forme de systèmes basés sur la recherche de candidats dans DBpedia et Wikidata, le calcul d'une similarité syntaxique et des votes majoritaires [1, 5, 7]. MTab4Wikidata [18] adopte la correspondance floue et la "recherche à deux cellules" pour améliorer la prise en charge des fautes d'orthographe et des ambiguïtés dans le contenu des tableaux. Ce système a remporté le premier prix des défis SemTab 2019 et SemTab 2020.

# 3 Système DAGOBAH SL 2021

DAGOBAH est un processus de bout en bout annotant des tables relationnelles avec des éléments d'un graphe de connaissances tel que Wikidata. Ce processus se compose de quatre étapes exécutées en séquence tel qu'illustré dans la Figure 1. Etant donné une table relationnelle en entrée, l'étape de pré-traitement détermine un ensemble de métadonnées de la table ainsi que les cibles de l'annotation (Section 3.1). Le module de recherche de candidats collecte ensuite des entités candidates dans le graphe de connaissances pour chaque cellule cible de la table (Section 3.2). Le module de notation préliminaire évalue chacun de ces candidats afin de déterminer un score de confiance (Section 3.3). Les étapes suivantes visent à générer les annotations CTA ainsi que les annotations CPA (Section 3.4). Enfin, les annotations précédentes sont mises à contribution pour générer les annotations CEA (Section 3.4).

## 3.1 Pré-traitement des données tabulaires

Dans des cas d'utilisation réels, l'annotation des tables se révèle complexe en partie à cause de l'absence d'informations préalables sur leur structure et leur contenu. Ainsi, leur pré-traitement peut faciliter leur annotation. C'est pourquoi DAGOBAH propose des méthodes de pré-traitement visant à générer des métadonnées sur les tables via quatre tâches principales : la détection d'orientation, l'extraction d'en-têtes, l'identification de colonne clé [2] et le typage primitif des colonnes. Le typage primitif permet de détecter des entités nommées (e.g. localisation, organisation, personne), des littéraux avec unités (e.g. distance, vitesse, température) ou des littéraux divers (e.g. email, URL, adresse IP) [2].

## 3.2 Recherche d'entités candidates

L'étape de pré-traitement (et plus particulièrement le typage primitif) permet d'identifier les colonnes d'une table éligibles à l'étape de recherche d'entités candidates. Soit $e_m$ une cellule d'une colonne éligible. L'étape de recherche d'entités candidates extrait un ensemble d'entités candidates pertinentes $\mathcal{E}_c(e_m)$ d'un graphe cible. Le service de recherche de candidats de DAGOBAH est basé sur ElasticSearch et supporte actuellement Wikidata et DBpedia pour lesquels des indexes ont été générés :

**Entités Wikidata.** Le service de recherche de candidats collecte les items et les propriétés ainsi que leurs labels et alias dans toutes les langues disponibles. Pour augmenter la couverture du service, les alias associés à chaque entité sont enrichis avec 11 propriétés supplémentaires telles que P2561 (name), P1705 (native label) ou P742 (pseudonym).

**Entités DBpedia.** Le service collecte les ressources en anglais ainsi que leurs labels dans toutes les langues disponibles. Pour augmenter la couverture, les labels sont enrichis avec les valeurs de 25 propriétés telles que abbreviation, birthName ou originalTitle. En complément, les labels et les alias de toutes les entités redirigées sont également inclus.

Nous faisons la moyenne des distances d'édition sur les caractères et sur les tokens [3] pour évaluer la similarité entre

2. Actuellement, seul l'identification d'une colonne clé unique est supportée par l'outil.
3. `https://github.com/seatgeek/thefuzz`

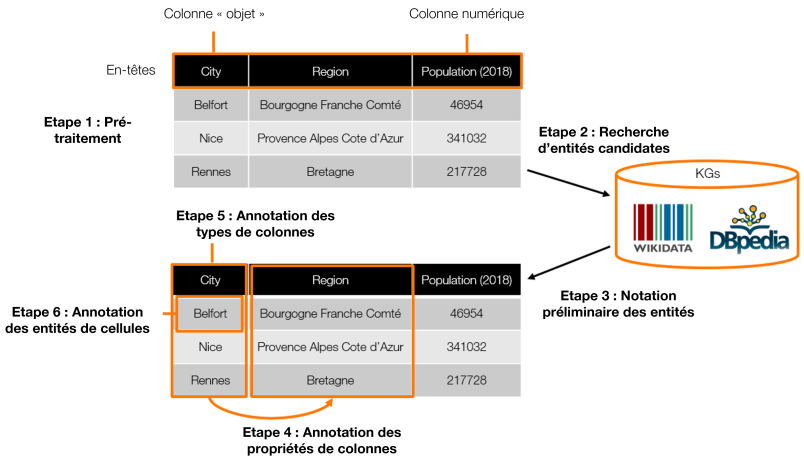



FIGURE 1 – Vue d'ensemble du processus d'annotation DAGOBAH.

une mention contenue dans une cellule et l'ensemble des labels de chaque entité candidate. Ce mode de fonctionnement permet de traiter au mieux les chaines de caractère présentant des ordonancements des sous-chaînes différents (e.g. "Elon Musk" et "Musk Elon").

### 3.3 Notation préliminaire des candidats

L'étape de notation préliminaire évalue la pertinence des entités candidates $e_c \in \mathcal{E}_c(e_m)$ d'une cellule $e_m$ à l'aide d'un score :

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

Ce score préliminaire est le produit d'un score de contexte et d'un score syntaxique $Sc_{sim}(e_c, e_m)$. Ce dernier facteur renvoie le plus haut ratio de correspondance, basé sur la distance de Levenshtein, entre la cellule et les labels et alias du candidat étudié. Les alias sont pénalisés avec un ratio pondéré par 0.9 car nous considérons que les labels ont plus d'importance pour la désambiguisation. Le facteur d'amplification $\gamma \in \mathbb{N}^+$ définit l'importance de la similarité syntaxique dans le calcul du score préliminaire. Nous avons déterminé, de manière empirique, que la valeur 2 était appropriée pour une utilisation du système sur les corpus du challenge SemTab2021.

Les améliorations du système DAGOBAH SL 2021 se concentre principalement sur le score de contexte, définit comme suit :

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

où $\mathcal{N}_{table}(e_m)$ est l'ensemble des cellules voisines de $e_m$ sur la même ligne et $\mathcal{N}_{graph}(e_c)$ est l'ensemble des nœuds voisins de l'entité $e_c$ dans le graphe de connaissances [4]. Pour chaque cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$, $sn_i$ est un score de correspondance défini par rapport à $\mathcal{N}_{graph}(e_c)$.

4. Les nœuds voisins sont connectés à $e_c$ via des chemins de prédicats dans le graphe de connaissances, quelque soit la direction des prédicats.

DAGOBAH SL 2021 résout deux problèmes de DAGOBAH SL 2020 inhérents au calcul du score de contexte :

**Evaluation coûteuse.** Chaque $sn_i$ était évalué en itérant sur l'ensemble des nœuds dans $\mathcal{N}_{graph}(e_c)$ pour trouver la meilleure correspondance. Par conséquent, un problème de performance survient lorsque l'algorithme doit noter une entité très générique du graphe de connaissances présentant des centaines voire des milliers de propriétés. Par exemple, considérons la cellule "Belfort" dans la Figure 1 et l'entité Wikidata candidate Q171545. Pour vérifier si la cellule "Bourgogne Franche Comté" est dans le contexte de Q171545, nous devons effectuer une comparaison avec chacun des $\sim$ 1000 nœuds de $\mathcal{N}_{graph}(Q171545)$ ce qui inclus Q142 (France), Q3371185 (Paul Faivre), etc. (Figure 2a).

**Contexte du graphe à un saut.** $\mathcal{N}_{graph}(e_c)$ est l'ensemble des nœuds situés à un saut de $e_c$ dans le graphe. Par conséquent, une cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ correspondant à un nœud situé à deux sauts de $e_c$ n'était pas prise en compte dans le contexte de $e_c$. Par exemple, soit le contexte à un saut de Q171545 (Belfort) dans la Figure 2a, nous considérions, à tort, que Bourgogne Franche Comté n'avait pas de relations avec Belfort bien qu'il s'agisse de la région du Territoire de Belfort dont la capitale est Belfort.

DAGOBAH SL 2021 améliore l'efficacité du calcul et l'expressivité du score de contexte en évitant une notation exhaustive et en exploitant des contextes d'entités plus expressifs via la considération de nœuds à deux sauts.

#### 3.3.1 Exploitation du contexte des entités du graphe de connaissances

Le score de correspondance du voisinage $sn_i$ défini dans l'Equation (2) indique si une cellule voisine $n_i$ de $e_m$ correspond à un nœud voisin de $e_c$. Le calcul de $sn_i$ peut se résumer à la recherche d'une entité candidate pour $n_i$

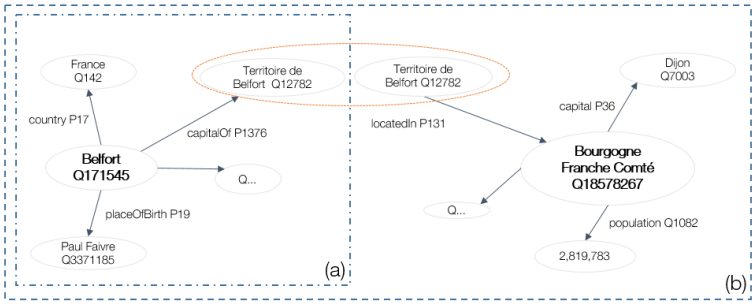



FIGURE 2 – Contexte (i.e. voisinage) de l'entité Q171545 (Belfort) dans le graphe Wikidata. (a) Contexte à un saut de Q171545. (b) Le contexte est étendu via l'intersection de sous-graphe.

dans $\mathcal{N}_{graph}(e_c)$ et à l'évaluation de sa similarité. Dans notre exemple précédent, Q18578267 est une entité candidate pour la cellule "Bourgogne Franche Comté" dans le contexte à deux sauts $\mathcal{N}_{graph}(Q171545)$ (Figure 2b). A partir de cette observation, nous proposons une méthode pour calculer efficacement le score $sn_i$. L'étape de recherche d'entités candidates (Section 3.2) génère des entités candidates $\mathcal{E}_c(e_m)$ pour une cellule cible $e_m$ mais également des entités candidates $\mathcal{E}_c(n_i)$ pour des cellules voisines $n_i$. Par conséquent, nous vérifions si une entité candidate $e_i \in \mathcal{E}_c(n_i)$ est dans $\mathcal{N}_{graph}(e_c)$. Dans ce cas, $sn_i$ est simplement calculé en comparant les labels de la cellule voisine $n_i$ et le nœud correspondant $e_i$. Ce point permet d'éviter des comparaisons additionnelles avec d'autres nœuds de $\mathcal{N}_{graph}(e_c)$.

Pour vérifier si $e_i \in \mathcal{E}_c(n_i)$ est dans $\mathcal{N}_{graph}(e_c)$, nous vérifions si $e_i$ est connecté à $e_c$ via un chemin de prédicats dans le graphe de connaissances. Le calcul de ces chemins est un élément important dans le calcul du score. Pour trouver efficacement un chemin de prédicats entre $e_c$ et $e_i$, nous extrayons les sous-graphes à un saut $\mathcal{G}_{e_c}$ et $\mathcal{G}_{e_i}$ de $e_c$ et $e_i$. Si un nœud intermédiaire $v$ est présent dans $\mathcal{G}_{e_c}$ et $\mathcal{G}_{e_i}$, les chemins pointant sur $v$ dans les deux sous-graphes sont concaténés. Dans notre exemple, le chemin de prédicats suivant a été identifié : Belfort $\xrightarrow{\text{capitalOf}}$ Territoire de Belfort $\xrightarrow{\text{locatedIn}}$ Bourgogne Franche Comté. Seuls les sous-graphes à un saut étant pris en compte, les chemins de prédicats résultant ont une longueur maximum de deux. Cette approche permet d'enrichir les informations sur une entité en incluant non seulement les voisins directs mais également les voisins indirects à une distance de deux sauts. Ces contextes enrichis du graphe permettent d'augmenter les chances de correspondance avec une cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ et rendent ainsi le score de contexte plus précis. Après des tests, nous faisons l'hypothèse que pour l'interprétation de tables avec Wikidata, des chemins de taille supérieure à deux sont peu significatifs et apportent du bruit pouvant impacter négativement la pertinence du score de contexte.

### 3.3.2 Notation souple du contexte

Dans l'Equation (2), les scores de correspondances du voisinage $sn_i$ sont pondérés pour calculer le score final d'une entité. En effet, chaque cellule voisine $n_i \in \mathcal{N}_{table}(e_m)$ contribue à un niveau différent à l'annotation de la cellule cible $e_m$ avec un poids $w_i$ défini par l'Equation (3) :

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

**(3a)** Les cellules contenant des entités devraient avoir une plus grande importance que les cellules contenant des littéraux (e.g. date, mesure avec ou sans unités, nombre) compte tenu du manque de méthodes de désambiguïsation des littéraux (e.g. normalisation des dates, détection des unités/normalisation/conversion). C'est la raison pour laquelle nous fixons la valeur de $se_i$ à 1.0 dans le cas où la cellule voisine $n_i$ contient une entité et à 0.15 si $n_i$ contient un littéral.

**(3b)** Une cellule voisine sur la partie gauche de la table a plus de chance d'être un contexte significatif pour la cellule cible. Par conséquent, $d(col_i)$ est la distance entre la colonne $col_i$ et la première colonne de type "entité" de la table.

**(3c)** Les cellules $n_i$ appartenant à une colonne voisine très connectée à la colonne cible devrait avoir un plus grand poids dans le contexte. Par conséquent, nous prenons en compte la connectivité $cnt(col_i)$ d'une cellule voisine par rapport à la colonne cible. La connectivité est définie ici comme le nombre d'occurrence de la propriété la plus souvent observée entre les deux colonnes.

**(3d)** Les nœuds voisins de l'entité candidate $e_c$ dans $\mathcal{N}_{graph}(e_c)$ peuvent fournir différents contenus informationnels étant donné que certains voisins peuvent être "sémantiquement plus proches" de $e_c$ que d'autres. Par exemple, si nous considérons le contexte à deux sauts de l'entités Q171545 (Belfort) présenté dans la Figure 3, Q18578267 (Bourgogne Franche Comté) est plus pertinent que Q30 (United States of America) car le chemin Belfort $\xrightarrow{\text{capitalOf}}$ Territoire de Belfort $\xrightarrow{\text{locatedIn}}$

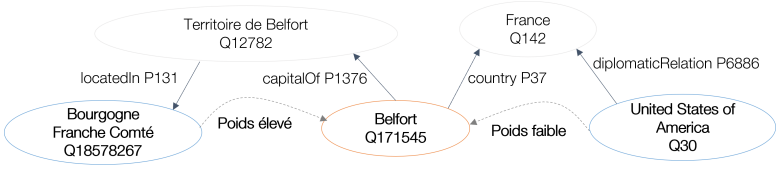



FIGURE 3 – Les nœuds voisins de Belfort (Q171545) contribuent de manière différente à son contenu informationnel.

Bourgogne Franche Comté porte davantage d'informations que le chemin Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America. La valeur de vérité $\tau(e_i)$ [6] d'un nœud voisin $e_i$ rend compte de cette différence en mesurant la capacité discriminative d'un chemin $\tau(e_c \xrightarrow{p_1} v \xrightarrow{p_2} e_i)$ et est défini comme suit :

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

où $g(v)$ est la généricité du nœud intermédiaire $v$, i.e. le nombre de ses relations entrantes et sortantes dans le graphe de connaissances. Il est à noter que les voisins directs (i.e. les chemins de longueur 1) obtiennent toujours une valeur de vérité de 1.0.

## 3.4 Tâches d'annotation

### 3.4.1 Annotation des propriétés de colonnes (CPA)

La tâche de CPA identifie la relation sémantique $r$ la plus adaptée pour une paire ordonnée de colonnes. Nous adoptons une stratégie de vote majoritaire reposant sur les occurrences et les scores de confiances cumulés des lignes pour $r$ (pour plus de détails, voir [11]). Il est à noter que, conformément à la Section 3.3, $r$ peut être un chemin de prédicats de longueur 1 (i.e., $\xrightarrow{p}$), un chemin unidirectionnel de longueur 2 (i.e. $\xrightarrow{p1}\xrightarrow{p2}$ ou $\xleftarrow{p1}\xleftarrow{p2}$) ou un chemin bidirectionnel de longueur 2 (i.e. $\xrightarrow{p1}\xleftarrow{p2}$ ou $\xleftarrow{p1}\xrightarrow{p2}$).

### 3.4.2 Annotation des types de colonnes (CTA)

La tâche de CTA a pour but d'identifier le type le plus représentatif et le plus spécifique d'une colonne donnée. Pour cela, les types des entités candidates de la colonne sont collectés et une stratégie de vote majoritaire est appliquée pour déterminer le type le plus précis (voir [11] pour plus de détails sur les méthodes d'enrichissement de type et les calculs de scores).

### 3.4.3 Annotation des entités de cellules (CEA)

La tâche de CEA sélectionne pour une cellule $e_m$ l'entité la plus pertinente parmi les entités candidates $e_c \in \mathcal{E}_c(e_m)$ collectées dans le graphe de connaissances. Cette étape s'appuie à la fois sur la notation préliminaire des entités et sur les informations fournies par le CTA et le CPA pour calculer la note finale des entités candidates. En effet, la notation préliminaire d'une entité candidate $e_c$ tient uniquement compte des informations locales, i.e. les informations de la ligne à laquelle elle appartient. La prise en compte du type de colonne fourni par le CTA et de la propriété identifiée par le CPA permet de prendre en compte des informations globales. Par conséquent, le score final $Sc(e_c, e_m)$ d'une entité candidate $e_c$ est calculé comme suit :

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

Si $e_c$ appartient au type généré par le CTA pour la colonne, alors $score_{CTA}$ est égal au score attribué à ce type et 0 dans le cas contraire. Via $\overline{score_{CPA}}$, nous calculons la moyenne des scores des relations identifiées par le CPA impliquant la colonne de $e_c$. Pour chaque relation, si $e_c$ appartient au domaine ou au co-domaine (selon l'orientation de la relation), alors nous considérons le score de cette relation, sinon, le score est fixé à 0. Pour renforcer (resp. affaiblir) un CTA/CPA fréquent (resp. peu fréquent) lors de la mise à jour de $Sc(e_c, e_m)$, un coefficient $\alpha$ (resp. $\beta$) est utilisé et défini par $\frac{occurrence(CTA)}{2}$ (resp. $\frac{\overline{occurrence(CPA)}}{2}$). Il est à noter que le nombre d'occurrences du CTA/CPA est divisé par 2 pour accorder davantage d'importance au score préliminaire $PSc(e_c, e_m)$.

# 4 Evaluation

## 4.1 Configurations

Afin d'évaluer l'apport des contextes de graphe à un et à deux sauts ainsi que de la notation de contexte souple définis dans la Section 3, nous définissons quatre configurations pour les expériences :

**Configuration 1** Le score de contexte d'une entité est calculé en utilisant uniquement le voisinage à un saut du graphe de connaissances. Les poids $w_i$ ne sont pas calculés à l'aide de l'Equation (3) mais fixés à 1.0 pour les entités et 0.15 pour les littéraux.

**Configuration 2** Le score de contexte d'une entité est calculé en utilisant le voisinage à deux sauts du graphe de connaissances. Les poids $w_i$ ne sont pas calculés à l'aide de l'Equation (3) mais sont fixés à 1.0 pour les voisins à un saut, 0.25 pour les voisins à deux sauts et 0.15 pour les littéraux.

**Configuration 3** Le score de contexte d'une entité est calculé en utilisant le voisinage à deux sauts du graphe de connaissances. Les poids $w_i$ sont calculés à l'aide de l'Equation (3). Cette configuration permet de tester si des contextes plus riches associés à une notation stricte permet de générer de meilleures annotations.

**Configuration 4** Ce paramétrage restreint la configuration 3 au voisinage à un saut et aux voisins liés par un chemin unidirectionnel de longueur 2 dans le graphe. Cette configuration permet d'évaluer l'impact des chemins bidirectionnels qui semblent être moins informatifs (et amenant parfois du bruit) mais utiles dans certains cas bien ciblés.

## 4.2 Résultats

### 4.2.1 Evaluation expérimentale

Les résultats pour les quatre configurations définies précédemment sont donnés dans la Table 1. Il est à noter que les performances de DAGOBAH se sont continuellement améliorées tout au long du challenge SemTab2021. Ainsi, les résultats de l'évaluation sont basés sur la dernière version de DAGOBAH mais nous indiquons également les résultats soumis lors des différentes phases du challenge dans les cellules grisées ainsi que le meilleur score parmi les participants du challenge [5], pour comparaison. Afin de valider la pertinence des modifications proposées dans les Sections 3.2 et 3.3, nous incluons également les scores du système DAGOBAH 2020 pour les tables du Round 1 annotées à l'aide de Wikidata. Les configurations des soumissions {1,2,3,4}* sont similaires aux configurations {1,2,3,4} définies précédemment avec quelques adaptations sur l'initialisation des scores et des poids. Cela n'impacte toutefois pas les scores de CEA mais a en revanche un impact sur les performances du CTA. En effet, le CTA est très sensible aux scores d'entités et aux poids attribués à la taxonomie pour la sélection du type le plus spécifique parmi l'ensemble des types possibles pour les entités (types directs, parents, etc.). DAGOBAH obtient d'excellents résultats sur les jeux de données synthétiques (Round 2) tandis que les jeux de données générés manuellement et présentant des dispositions plus complexes semblent être traités de manière moins satisfaisante (Rounds 1 et 3). Sur le corpus HardTable, l'utilisation de contextes plus riches et de la technique de notation souple ne semble pas amener de gain. Cela peut s'expliquer par le fait que les tableaux de ce corpus sont presque entièrement représentés dans le graphe de connaissances et que les colonnes peuvent donc être désambiguisées seulement à partir de leur contenu. A l'inverse, le corpus BioTable contient des ambiguités plus complexes avec des chevauchements de contenu entre les colonnes empêchant leur désambiguisation (e.g. la colonne "Gene" peut être confondue avec la colonne "Protein", les valeurs étant souvent similaires). L'annotation semble donc bénéficier de contextes de graphes plus riches. Pour BioDivTable, la configuration 4 est celle obtenant les scores les plus bas, tandis que la configuration 1 est comparable à la configuration 3. Nous supposons que les chemins unidirectionnels de longueur 2 apportent du bruit pouvant expliquer les faibles performances de la configuration 4.

En règle général, les configurations 2, 3 et 4 sont plus précises pour le CEA que la configuration 1. La récupération du contexte de graphe à deux sauts semble donc être un ajout bénéfique permettant de récupérer des informations pertinentes. Les meilleures performances des configurations 3 et 4 vis à vis de la configuration 2 montre l'efficacité de la notation de contexte souple. Nous notons que la configuration 3 atteint des performances proches de la configuration 4. Ainsi, les chemins unidirectionnels (i.e. $\xrightarrow{p_1}\xrightarrow{p_2}$ et $\xleftarrow{p_1}\xleftarrow{p_2}$) apportent suffisamment d'informations et permettent d'obtenir des résultats équivalents par rapport à la configuration considérant à la fois les chemins unidirectionnels et bidirectionnels. De plus, l'influence négative du bruit apporté par les chemins bidirectionnels (e.g. Belfort $\xrightarrow{\text{country}}$ France $\xleftarrow{\text{diplomaticRelation}}$ United States of America) est limitée par le calcul de score de contexte souple qui évite une dégradation de la qualité de l'annotation. Cela permet aux chemins bidirectionnels pertinents de contribuer positivement au score de l'entité. On peut observer que les performances du CTA et du CPA ne sont pas aussi élevées qu'envisagé sur la plupart des corpus, et ce, malgré de bonnes performances de CEA. Le développement de stratégies plus performantes pour la sélection du type et des relations fera l'objet de travaux futurs.

### 4.2.2 Corpus BioDivTab et GitTables

Il est à noter que pour les corpus BioDivTab et GitTables, nous avons adapté les algorithmes DAGOBAH présentés dans cet article. En effet, pour le corpus BioDivTab, les types primitifs générés par le pré-traitement ont été utilisés pour discriminer les colonnes "entités" et les colonnes contenant des littéraux. Une colonne contient des littéraux si elle contient des valeurs numériques, des dates, des unités ou des valeurs diverses. Sinon, la colonne est considérée comme une colonne d'entités et ses mentions peuvent donc être utilisées par le module de recherche de candidats. Pour le corpus GitTables, des règles de correspondance ont été définies entre les types primitifs et des classes de Schema.org et de l'ontologie DBpedia.

# 5 L'interprétation de données tabulaires à Orange

Pour améliorer la pertinence de DAGOBAH sur des cas d'utilisation industriels réels, nous avons adopté une approche Test & Learn. Dans cette optique, les algorithmes de DAGOBAH sont mis à disposition au sein de l'entreprise pour permettre aux collaborateurs internes de tester les outils d'annotation. Cette mise à disposition s'effectue via deux vecteurs : une API REST nommée `TableAnnotation` et une interface graphique nommée DAGOBAH UI.

## 5.1 API TableAnnotation

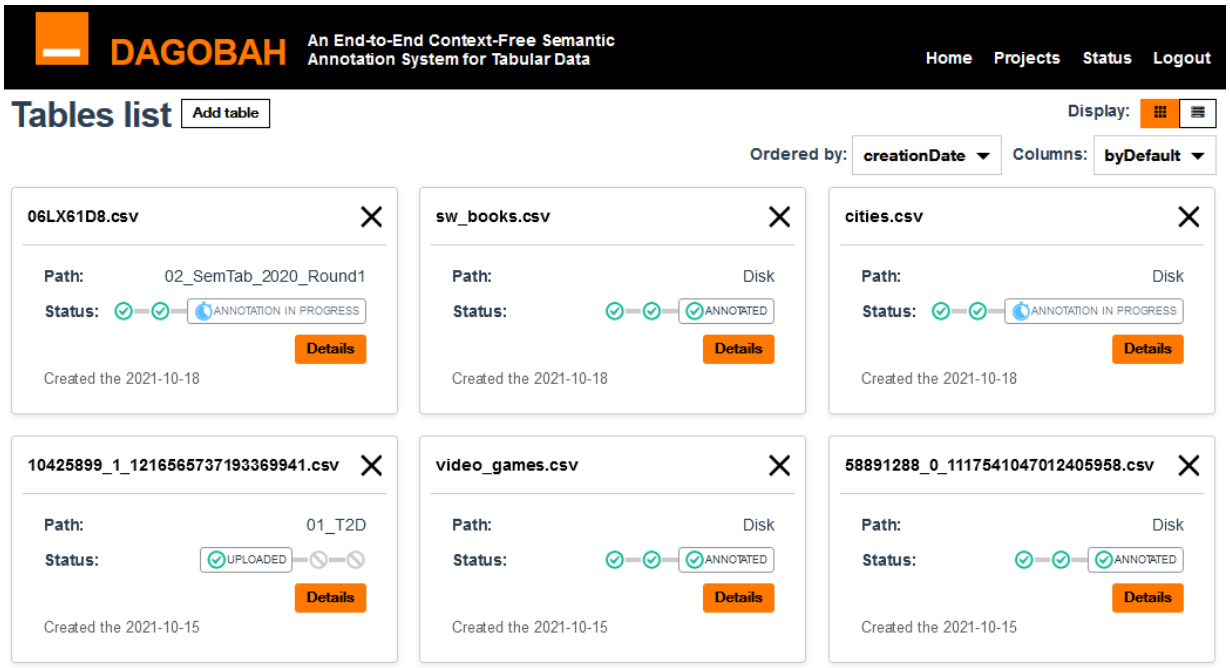
Cette API REST est déployée sur le portail Orange Developer. [6] Elle fournit des services de pré-traitement de données tabulaires, d'annotation sémantique et également de recherche d'entités candidates permettant, à partir d'une

5. Les résultats complets sont disponibles en ligne : `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/index.html#results`

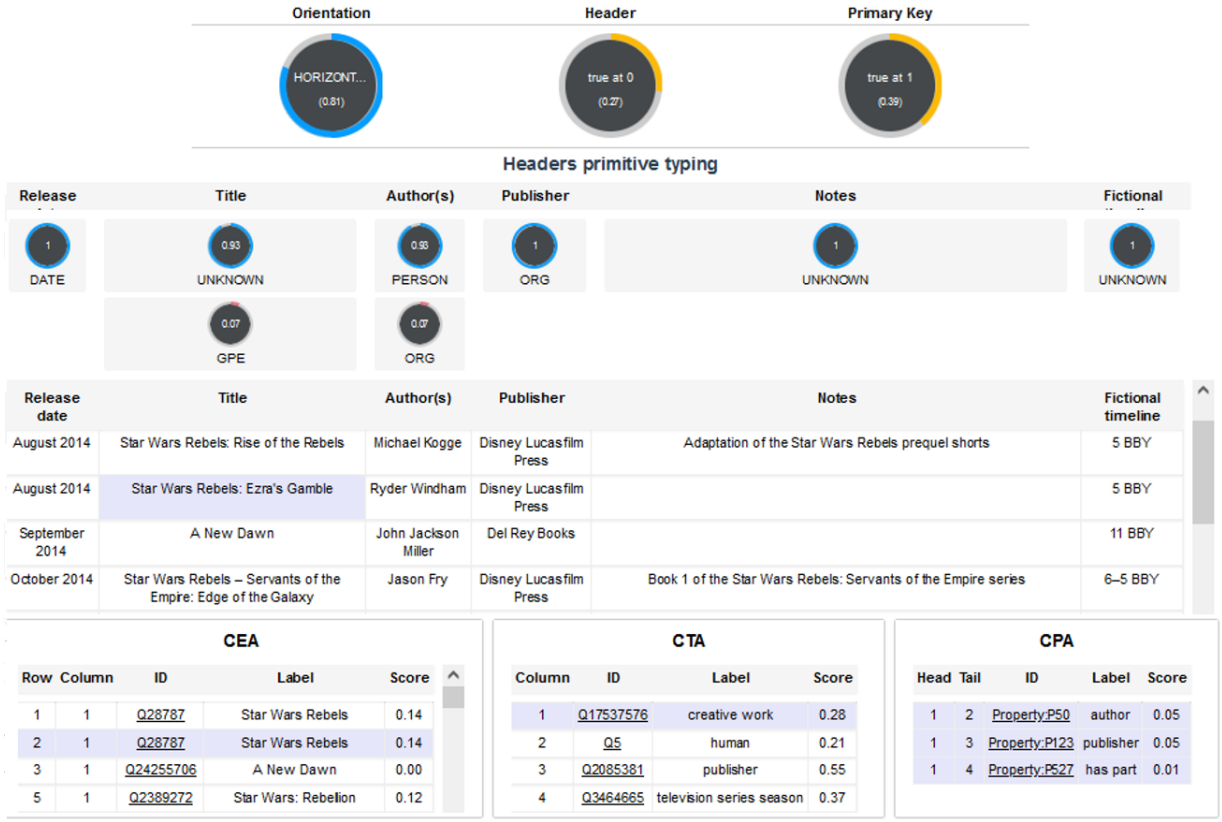
6. `https://developer.orange.com`

TABLE 1 – Comparaison des configurations expérimentales et des performances du système DAGOBAH sur les Rounds 1, 2, et 3 du challenge SemTab2021 (dans les cellules grisées). "F1" signifie F1-score, "P" signifie Precision. Les meilleurs résultats sont mis en valeur en gras.

| Corpus | Configuration | CTA | | CEA | | CPA | |
|---|---|---|---|---|---|---|---|
| | | F1 | P | F1 | P | F1 | P |
| Round 1 – WDTable | Configuration 1 | 0.793 | 0.793 | 0.913 | 0.913 | - | - |
| | Configuration 2 | 0.790 | 0.790 | 0.923 | 0.923 | - | - |
| | Configuration 3 | 0.783 | 0.783 | **0.926** | **0.926** | - | - |
| | Configuration 4 | 0.783 | 0.783 | 0.924 | 0.924 | - | - |
| | DAGOBAH 2020 | 0.743 | 0.743 | 0.830 | 0.841 | - | - |
| | Configuration 2* | **0.832** | **0.832** | 0.923 | 0.923 | - | - |
| | Top 1 SemTab2021 | 0.728 | 0.73 | 0.907 | 0.907 | - | - |
| Round 1 – DBPTable | Configuration 1 | 0.25 | 0.25 | 0.935 | 0.935 | - | - |
| | Configuration 2 | 0.27 | 0.27 | 0.946 | 0.946 | - | - |
| | Configuration 3 | 0.274 | 0.274 | **0.947** | **0.947** | - | - |
| | Configuration 4 | 0.274 | 0.274 | **0.947** | **0.947** | - | - |
| | Configuration 2* | 0.422 | 0.424 | 0.945 | 0.946 | - | - |
| | Top 1 SemTab2021 | **0.46** | **0.468** | 0.692 | 0.692 | - | - |
| Round 2 – BioTable | Configuration 1 | 0.874 | 0.874 | 0.882 | 0.882 | 0.898 | **0.901** |
| | Configuration 2 | 0.911 | 0.911 | 0.916 | 0.916 | **0.899** | 0.899 |
| | Configuration 3 | 0.915 | 0.915 | 0.950 | 0.951 | **0.899** | 0.899 |
| | Configuration 4 | 0.916 | 0.916 | **0.970** | **0.970** | **0.899** | 0.899 |
| | Configuration 4* | 0.916 | 0.916 | 0.970 | 0.970 | 0.899 | 0.899 |
| | Top 1 SemTab2021 | **0.956** | **1** | 0.964 | 0.964 | 0.899 | 0.899 |
| Round 2 – HardTable | Configuration 1 | 0.968 | 0.969 | 0.975 | 0.976 | 0.996 | 0.997 |
| | Configuration 2 | 0.968 | 0.969 | 0.976 | 0.976 | 0.996 | 0.997 |
| | Configuration 3 | 0.968 | 0.969 | 0.976 | 0.976 | 0.996 | 0.997 |
| | Configuration 4 | 0.968 | 0.968 | 0.976 | 0.976 | 0.996 | 0.997 |
| | Configuration 3* | 0.976 | 0.976 | 0.975 | 0.976 | 0.996 | 0.996 |
| | Top 1 SemTab2021 | **0.977** | **0.977** | **0.985** | **0.985** | **0.997** | **0.998** |
| Round 3 – BioDivTable | Configuration 1 | 0.338 | 0.339 | 0.619 | 0.64 | - | - |
| | Configuration 2 | 0.335 | 0.335 | 0.60 | 0.62 | - | - |
| | Configuration 3 | 0.344 | 0.345 | **0.62** | **0.641** | - | - |
| | Configuration 4 | 0.343 | 0.343 | 0.475 | 0.491 | - | - |
| | Configuration 4* | 0.381 | 0.382 | 0.496 | 0.497 | - | - |
| | Top 1 SemTab2021 | **0.593** | **0.595** | 0.602 | 0.611 | 0.947 | 1 |
| Round 3 – HardTable | Configuration 3* | **0.99** | **0.99** | **0.974** | **0.974** | 0.991 | **0.995** |
| | Top 1 SemTab2021 | 0.984 | 0.984 | 0.968 | 0.968 | **0.993** | 0.994 |
| Round 3 – GitTables DBP | Pré-traitement + Mapping | **0.07** | **0.117** | - | - | | - |
| | Top 1 SemTab2021 | 0.041 | 0.042 | - | - | - | - |
| Round 3 – GitTables SCH | Pré-traitement + Mapping | 0.183 | 0.185 | - | - | - | - |
| | Top 1 SemTab2021 | **0.205** | **0.943** | - | - | - | - |




(a) DAGOBAH UI permet de créer des projets et de charger des tables à partir du système de fichier local ou à partir de corpus de référence (e.g. T2D, SemTab).




(b) DAGOBAH UI permet d'afficher les résultats générés par les outils de pré-traitement et d'annotation. En partie haute, l'outil affiche les informations de pré-traitement (e.g. orientation, en-têtes) ainsi que la table nettoyée. En partie basse, une vue interactive permet à l'utilisateur de naviguer dans les annotations CEA, CTA et CPA.

FIGURE 4 – Fonctionnalités de DAGOBAH UI.

mention, de collecter des entités Wikidata ou DBpedia potentiellement correspondantes. Cette API est accessible à l'ensemble des collaborateurs des entités R&D du groupe ainsi que des unités d'affaires, sur invitation. Nous planifitions d'ouvrir plus largement l'accès à cette API dans un futur proche.

### 5.2 DAGOBAH UI

Cette interface graphique permet à des collaborateurs non familiers avec le développement ou l'intelligence artificielle d'utiliser les fonctions de l'API `TableAnnotation` sur leurs tables et de visualiser les résultats sous une forme intelligible et ergonomique. Les utilisateurs ont la possibilité de charger des tables dans leurs projets d'annotation (Figure 4a) puis de lancer le pré-traitement de ces dernières ainsi que l'annotation sémantique. Les résultats de ces processus peuvent ensuite être visualisés (Figure 4b). DAGOBAH UI est un outil très puissant pour démontrer la valeur des techniques d'interprétation automatique de tables au sein du groupe Orange mais également auprès de prospects externes. Une vidéo de démonstration de l'interface graphique est disponible à `https://tinyurl.com/dagobah-ui`. Les développements récents sur cette interface permettent *(i)* l'enrichissement de graphes de connaissances à partir d'éléments de la table non présents dans le graphe, *(ii)* l'enrichissement de la table à partir du graphe de connaissances afin de compléter des valeurs manquantes ou d'ajouter de nouvelles colonnes, et *(iii)* la visualisation interactive du graphe de connaissances cible avec une mise en valeur des annotations résultants des étapes de CEA, CTA, CPA ainsi que des nouveaux triplets générés à partir de la table.
Cette interface utilisateur permet aux collaborateurs de saisir l'interêt de l'annotation sémantique pour des cas d'utilisation industriels. Inversement, l'équipe de recherche DAGOBAH peut identifier les défis associés à ces cas d'utilisation, ce qui constitue un apport précieux pour la feuille de route du projet. Bien que le déploiement et l'adoption des méthodes de STI chez Orange n'en soient qu'à leurs débuts, des tests sur différents cas d'utilisation ont lieu depuis plus d'un an via l'API `TableAnnotation` qui a répondu à plus de 200 000 requêtes. Plusieurs domaines sont envisagés comme cibles prioritaires pour l'annotation sémantique, incluant le divertissement (e.g. annotation de catalogues de films), la gouvernance des données ou la santé.

## 6 Discussion

Les corpus de données proposés cette année par le challenge SemTab2021 ont permis de prendre en compte une plus grande variété de problématiques associées à l'interprétation automatique de données tabulaires. Cette édition a notamment intégré de nouveaux domaines de connaissances (e.g. biomédical et données Git) et a ajouté de nouvelles contraintes sur l'annotation avec le support de graphes de connaissances multiples (Wikidata et DBpedia) et l'annotation à l'aide de schémas uniquement (Schema.org et l'ontologie DBpedia). Ces challenges nous ont permis d'améliorer les stratégies d'annotation du système DAGOBAH avec notamment l'exploitation des types primitifs générés par le pré-traitement et l'utilisation de contextes de graphes enrichis.

Néanmoins, de nouvelles directions de recherche peuvent encore être explorées pour faire face à l'hétérogénéité des types de tableaux publiés sur le Web. Ainsi, la structure des tableaux et les relations internes pourraient être prises en compte (e.g. orientation des tableaux, cellules imbriquées, concaténation pour mise en page, cellules à valeurs multiples, sujets répartis dans plusieurs colonnes comme les noms et prénoms d'une personne, etc). De plus, une problématique demeure dans le traitement des données hors graphe de connaissances, i.e. des entités non présentes dans un graphe cible donné, ce qui est souvent le cas pour des données spécifiques aux entreprises. Il convient de noter que les données hors graphe de connaissances ont commencé à être abordées avec le corpus GitTables. Ce dernier nécessitait en effet d'annoter des tables avec Schema.org et l'ontologie DBpedia uniquement. Cependant, cette tâche n'était pas entièrement conforme avec la définition du CTA utilisée par la communauté car les annotations recherchées mélangeaient des classes et des propriétés. Ces annotations hétérogènes peuvent conduire à des évaluations incohérentes. Au delà des données hors graphe, il serait intéressant d'évaluer la portabilité de l'approche à des graphes de domaines (e.g. biomédicaux, linguistiques) et leur apport pour l'annotation de jeux de données spécifiques.

Nous avons récemment proposé une classification reflétant l'hétérogénéité des tables que l'on peut rencontrer ainsi qu'un inventaire exhaustif des méthodes, à base de règles et d'heuristiques, ou à base d'apprentissage profond pour l'interprétation sémantique de données tabulaires [16]. Nos travaux en cours se concentrent sur l'interprétation de tables où une grande part des mentions d'une colonne ne trouve pas de correspondance, ou avec peu de lignes et donc peu de contexte. Dans ces cas difficiles, nous cherchons à tirer profit des modèles de langage et des méthodes de plongement de graphes qui pourraient apporter un complément intéressant aux stratégies de calcul de score de contexte.

## 7 Conclusion

Dans cet article, nous avons présenté les améliorations apportées au système DAGOBAH [3]. Grace à un mécanisme de recherche de candidats optimisé, l'enrichissement des contextes du graphe et la notation souple, DAGOBAH a obtenu la meilleure performance lors du challenge SemTab2021.[7] Les travaux futurs auront pour objectif d'augmenter la précision de l'annotation sur des tables présentant des mentions très ambiguës. Nous avons notamment l'ambition d'exploiter des dictionnaires fournissant des abréviations ou des acronymes. Pour assurer la généricité de notre approche, de tels dictionnaires devraient être construits à partir de grandes quantités de documents et être applicables à divers ensembles de données.

7. `https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2021/index.html#results`

## Remerciements

Les auteurs remercient Christophe Sarthou-Camy et Guillaume Jourdain pour leurs contributions importantes dans le développement de DAGOBAH UI.

## Références

[1] Nora Abdelmageed and Sirko Schindler. JenTab Meets SemTab 2021's New Challenges. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab). CEURWS. org*, 2021.

[2] Yoan Chabot, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. DAGOBAH : An End-to-End Context-Free Tabular Data Semantic Annotation System. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 2553 of *CEUR Workshop Proceedings*, pages 41–48, 2019.

[3] Yoan Chabot, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. DAGOBAH : Un système d'annotation sémantique de données tabulaires indépendant du contexte. In *31$^{es}$ Journées francophones d'Ingénierie des Connaissances (IC)*, Angers, France, 2020.

[4] Yoan Chabot, Pierre Monnin, Frédéric Deuzé, Viet-Phi Huynh, Thomas Labbé, Jixiong Liu, and Raphaël Troncy. A Framework for Automatically Interpreting Tabular Data at Orange. In *20$^{th}$ International Semantic Web Conference (ISWC), Posters, Demos and Industry Tracks*, volume 2980 of *CEUR Workshop Proceedings*, 2021.

[5] Shuang Chen, Alperen Karaoglu, Carina Negreanu, Tingting Ma, Jin-Ge Yao, Jack Williams, Andy Gordon, and Chin-Yew Lin. Linkingpark : An integrated approach for semantic table interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2020.

[6] Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6), 2015.

[7] Marco Cremaschi, Roberto Avogadro, Andrea Barazzetti, and David Chieregato. MantisTable SE : an Efficient Approach for the Semantic Table Interpretation. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2020.

[8] Vincenzo Cutrona, Jiaoyan Chen, Vasilis Efthymiou, Oktie Hassanzadeh, Ernesto Jiménez-Ruiz, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, Madelon Hulsebos, Daniela Oliveira10, et al. Results of SemTab 2021. In *CEUR Workshop Proceedings*, 2021.

[9] Vasilis Efthymiou, Oktie Hassanzadeh, Mariano Rodriguez-Muro, and Vassilis Christophides. Matching web tables with knowledge base entities : From entity lookups to entity embeddings. In *16$^{th}$ International Semantic Web Conference (ISWC)*, pages 260–277, 2017.

[10] Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Frédéric Deuzé, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. DAGOBAH : Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 3103 of *CEUR Workshop Proceedings*, pages 19–31, 2021.

[11] Viet-Phi Huynh, Jixiong Liu, Yoan Chabot, Thomas Labbé, Pierre Monnin, and Raphaël Troncy. DAGOBAH : enhanced scoring algorithms for scalable annotations of tabular data. In *International Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, volume 2775 of *CEUR Workshop Proceedings*, pages 27–39, 2020.

[12] Yusra Ibrahim, Mirek Riedewald, and Gerhard Weikum. Making sense of entities and quantities in Web tables. In *International Conference on Information and Knowledge Management (CIKM)*, pages 1703–1712, 2016.

[13] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, and Kavitha Srinivas. SemTab 2019 : Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems. In *European Semantic Web Conference (ESWC)*, pages 514–530. Springer, 2020.

[14] Ernesto Jiménez-Ruiz, Oktie Hassanzadeh, Vasilis Efthymiou, Jiaoyan Chen, Kavitha Srinivas, and Vincenzo Cutrona. Results of SemTab 2020. In *CEUR Workshop Proceedings*, volume 2775, pages 1–8, 2020.

[15] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti. Annotating and searching web tables using entities, types and relationships. In *36$^{th}$ International Conference on Very Large Data Bases (VLDB)*, pages 1338–1347, 2010.

[16] Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. From Tabular Data to Knowledge Graphs : A Survey of Semantic Table Interpretation Tasks and Methods. *To appear in the Journal of Web Semantics*, 2022.

[17] Varish Mulwad, Tim Finin, Zareen Syed, and Anupam Joshi. Using linked data to interpret tables. In *1$^{st}$ International Workshop on Consuming Linked Data (COLD)*, 2010.

[18] Phuc Nguyen, Ikuya Yamada, Natthawut Kertkeidkachorn, Ryutaro Ichise, and Hideaki Takeda. SemTab 2021 : Tabular Data Annotation with MTab Tool. In *Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab)*, 2021.

[19] Chenwei Ran, Wei Shen, Jianyong Wang, and Xuan Zhu. Domain-specific knowledge base enrichment using wikipedia tables. In *IEEE International Conference on Data Mining (ICDM)*, pages 349–358, 2016.