Des données tabulaires à RDF : l'extraction de données de Google Fusion Tables ⋆

Gianluca Quercini, Chantal Reynaud-Delaître

Laboratoire de Recherche en Informatique, Université Paris Sud XI
{gianluca.quercini, chantal.reynaud}@lri.fr

Résumé : Durant ces dernières années, un volume très important de données RDF, respectant les principes des données liées (LOD) établis par Tim Berners-Lee, ont été publiées. Toutefois, ces données ne représentent qu'une petite partie des données du web déjà disponibles dans d'autres formats, des données qu'il est très difficile de traduire manuellement en RDF étant donné leur volume. Dans ce papier, nous présentons une approche d'extraction d'une partie de ces données via *Google Fusion Tables* dans le but de peupler une ontologie d'un domaine donné. Les données collectées sont issues de tableaux conçus par des millions d'internautes. Notre objectif est d'en extraire un sous-ensemble pertinent par rapport à des classes d'une ontologie que nous souhaitons instancier. Ces données seront ensuite représentées en RDF. Une évaluation partielle de notre approche a été effectuée sur des données relatives à des restaurants et des musées.

Mots-clés : Annotation sémantique, Données liées (Linked Data), Peuplement d'ontologie.

1 Introduction

Le *web sémantique* s'est donné comme objectif de dépasser les limites du web actuel et permet, entre autres, d'améliorer l'efficacité de la recherche d'information. Les moteurs de recherche actuels sont capables de retourner tout un ensemble de pages web pouvant contenir la réponse à une requête utilisateur (Experian Hitwise US (2010)), mais c'est à l'utilisateur d'étudier chacune de ces pages pour y trouver la réponse recherchée, celle-ci pouvant parfois être obtenue après recoupement d'informations contenues dans plusieurs pages. En effet, les moteur de recherche appliquent en-

⋆. Cette recherche est financée par l'Agence Nationale de la Recherche à travers le projet DataBridges (ANR-11 EITS 003 05)

core aujourd'hui très souvent des approches basées sur la reconnaissance de mots-clés et l'exploitation d'informations décrivant le contenu sémantique des pages n'en est encore qu'à ses débuts : *Knowledge Graph* [1] de Google en est un exemple (cf. le Bourlout (2012)).

En 2006, Berners-Lee (2006) a publié quatre règles, très simples, connues sous la dénomination *Principes des données liées*, pour aider à publier du contenu dans le cadre du web sémantique. Ceci a conduit à la création du projet *Linking Open Data*, qui est la concrétisation la plus visible du web sémantique [2] . Ce projet est par nature ouvert. Les données publiées correspondent à un nombre impressionnant d'ensembles de données RDF provenant d'individus ou d'organisations variées telles que la *BBC*, le *gouvernement américain* ou la *bibliothèque du Congrès*. Les données sont très diversifiées. On y trouve des données sur des lieux géographiques, des livres, des personnes, des médicaments, de la musique, des programmes de télévision ou de radio, et ces ensembles de données sont liés les uns aux autres. Toutes ces données interconnectées dans le cadre du projet *Linking Open Data* correspondent à ce qu'on appelle aujourd'hui le *LOD*, qui correspond *de facto* au site web officiel des données.

Actuellement, le *web des données* contient pratiquement 31 milliards de triplets RDF [2] . Il a atteint une masse critique de données disponibles et librement réutilisables pouvant être exploitées par des machines pour de nouveaux services liés aux évolutions des usages et devenant possibles par la publication des données publiques. Bien qu'énorme, le Web des données ne représente toutefois encore qu'une partie des données disponibles dans les pages du web, les données sur les cinémas ou l'actualité en sont ainsi deux exemples d'ensembles de données absentes. Comme le souligne Coetzee *et al.* (2008), les individus et les organisations seront d'autant plus prêts à rendre leurs données disponibles pour des applications du web sémantique, qu'ils seront en mesure de voir les avantages apportés à la gestion des données et à la recherche de réponses aux requêtes posées sur le web. Toutefois, pour cela, il est nécessaire que les applications du web sémantique disposent d'encore plus de données. Ceci montre l'urgence d'enrichir le web des données en mettant en place des mécanismes d'extraction et d'annotation automatique des données à partir du web des pages, à la façon de DBpedia (Bizer *et al.* (2009)) et Yago (Suchanek *et al.* (2007))

1. `http://www.google.com/insidesearch/features/search/knowledge.html`

2. `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`

mais à plus grande échelle.

Cafarella *et al.* (2008) ont montré qu'une grande partie des données du web sont représentées dans des tableaux. Du point de vue extraction de données, les tableaux ont au moins trois atouts par rapport au texte :

1. Leur structure est régulière. Chaque cellule contient une chaîne de caractères alpha-numérique représentant une information atomique porteuse de sens. En revanche, un texte doit être analysé de façon à repérer les suites de mots ou expressions conformes à certains modèles.
2. Ils mettent en avant l'information la plus importante (ex. le nom, la localisation, le prix) caractérisant des entités (ex. des restaurants), alors qu'un texte non structuré citera ces éléments au milieu d'autres sans les distinguer plus particulièrement.
3. Ils fournissent des informations, par exemple des éléments de nature statistique, qu'on ne trouve généralement pas dans du texte.

En revanche, les tableaux ont certains inconvénients par rapport au texte, le plus important étant celui de l'absence du contexte. Ceci rend en particulier très difficile la tâche d'annotation.

A la différence de Cafarella *et al.* (2008), notre objectif n'est pas d'identifier des tableaux dans des pages Web. Nous exploitons l'ensemble de tableaux mis à disposition dans *Google Fusion Tables* (GFT), une application web très intéressante permettant de partager des données tabulaires via Internet. Il n'y a pas de statistiques officielles sur le nombre de tables dans GFT. Toutefois, Google met à disposition un moteur de recherche pour chercher les tables dans GFT et dans les sites web. Nous avons fait plusieurs requêtes pour estimer le nombre de tables dans GFT. Le résultat démontre que Google indexe à l'heure actuelle plus de 20 millions de tables. Dans ce papier, nous présentons une approche ayant pour but d'extraire des données des tableaux GFT et de les traduire en RDF. Il s'agit de :

- **Identifier** des tableaux de GFT contenant des données de ce domaine d'intérêt,
- **Extraire** de l'information des tableaux préalablement sélectionnés en utilisant l'ontologie,
- **Peupler** l'ontologie avec les données extraites.

Cet article est structuré de la façon suivante. La section 2 est un bref état de l'art. En section 3, nous présentons le contexte du problème, c'est-à-dire l'ontologie qui est exploitée et *Google Fusion Tables*. La section 4

est consacrée à la description de notre approche. Nous y décrivons le processus d'identification des tableaux et d'extraction, ce qui permet de caractériser les données obtenues de GFT pour peupler l'ontologie. La section 5 présente quelques résultats expérimentaux obtenus par application de notre approche sur des données de GFT portant sur des restaurants et des musées. Cet exemple, concernant des points d'intérêts (POIs) de la ville, concerne des données qui nous sont utiles dans le cadre du projet *ANR DataBridges*, dont l'objectif est d'intégrer des données en exploitant les technologies et les modèles des données liées, pour des applications en lien avec les villes numériques.

2 Travaux proches

Pour extraire des données d'un tableau, il faut comprendre la signification de l'information qui y est contenue. Ce problème a été largement étudié.

Amin *et al.* (2010) ont proposé un algorithme qui étiquette chaque colonne d'un tableau à l'aide d'une catégorie de *Wikipédia* décrivant au mieux son contenu. La catégorie retenue dépend des entités reconnues dans la colonne ainsi que de la catégorie étiquetant la colonne précédente.

L'algorithme proposé par Limaye *et al.* (2010) annote un tableau en typant les colonnes (un ou plusieurs types par colonne), en désignant par une relation binaire chaque couple de colonnes et en affectant un identifiant à chaque cellule. Les types, les relations et les identifiants proviennent de *Yago*, une ontologie construite au dessus de *Wikipédia* et décrite par Suchanek *et al.* (2007). Une annotation d'une table est un ensemble de variables aléatoires dont la distribution conjointe est représentée par un *modèle probabiliste graphique*. L'algorithme recherche les valeurs des variables maximisant leur probabilité conjointe. Venetis *et al.* (2011) adopte une approche similaire basée sur le modèle statistique du *maximum de vraisemblance*. Un tableau est annoté avec des entités, des types et des relations d'un catalogue obtenu par extraction d'informations de pages web, plus riche que *Yago*. Cette approche permet ainsi d'obtenir de meilleures annotations. Le problème, toutefois, de ces approches est leur incapacité à annoter des tableaux contenant des entités non répertoriées dans les catalogues utilisés. L'algorithme que nous proposons n'est pas basé sur un catalogue ; il peut donc identifier des entités non répertoriées.

Un travail très proche du nôtre est celui décrit par Guo *et al.* (2011), qui propose un outil appelé *ITEM* pour enrichir des bases de connais-

sances RDF avec des entités extraites de GFT. Dans un premier temps, les tableaux dont la structure correspond à la base de connaissances RDF sont extraits, puis les tuples qui ne sont pas dans la base y sont rajoutés. La différence essentielle avec notre travail porte sur le fait que nous ne supposons pas l'existence d'une base de connaissances RDF pour aider à identifier de nouvelles entités dans des tableaux GFT. Notre algorithme est suffisamment général pour retrouver des références à des entités dans des tables dont le schéma ne correspond pas forcément à celui d'une base de connaissances.

3 Description du contexte

L'approche que nous proposons a pour but de peupler une ontologie ; cette ontologie est utilisée pour guider le processus d'identification et d'extraction des données. Nous la présentons puis nous décrivons *Google Fusion Tables*.

3.1 L'ontologie

L'ontologie à peupler est une ontologie OWL. Dans l'exemple utilisé dans cet article, elle définit les points d'intérêts les plus importants d'une ville, ses musées, ses restaurants, ses théâtres, ses stades, ses hôtels et divers sites touristiques comme les monuments, tours et parcs nationaux (cf. Fig. 1).

Chaque classe est définie par des attributs qui ne sont pas représentés sur la figure 1 pour plus de clarté. Par exemple, les attributs les plus importants caractérisant les restaurants sont le nom, l'adresse, le prix moyen, le numéro de téléphone, l'adresse du site web et la catégorie. Cette ontologie comprend également des classes d'entités (ex. collection, artiste) en relation avec les points d'intérêt de la ville (ex. musée). Ces informations sont utilisées pour l'extraction des données des tableaux sur les restaurants. Le travail sur cette ontologie n'est pas encore totalement finalisé. La figure 1 ne correspond donc pas à la version finale qui pourra, in fine, être plus complète.

3.2 Google Fusion Tables

Google Fusion Tables est un service d'intégration et de gestion de données basé sur le *cloud* fourni par *Google*, qui permet de partager des

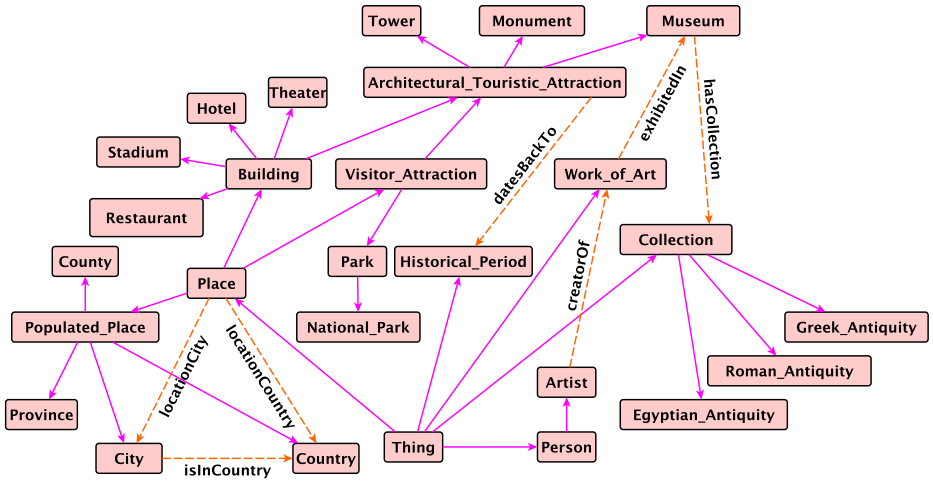

FIGURE 1 – Notre ontologie.

données en créant des tableaux via une interface intuitive facile à utiliser (Gonzalez *et al.* (2010)). Les tableaux peuvent être créés de toute pièce dans GFT ou provenir de *tableurs*, de *Google Docs* ou correspondre à des fichiers CSV importés ensuite dans GFT.

Le but déclaré de GFT est de permettre à ceux qui ont des données de les gérer alors qu'ils n'ont pas d'expertise en base de données (Zicari (2011)). La facilité d'utilisation de GFT devrait encourager le partage de données structurées, facilitant ainsi leur extraction et leur annotation.

Comme les tables d'une base de données relationnelle, chaque tableau a un nom unique qui, dans GFT est une chaîne de caractères alpha-numérique affectée automatiquement à sa création. Tous les tableaux de GFT sont accessibles via un navigateur web à l'adresse suivante : `http://www.google.com/fusiontables/DataSource?docid=[tableID]`, où *tableID* est l'identifiant GFT du tableau. Chaque colonne a un *nom* (ou *entête*) et un *type* (TEXT, NUMBER, LOCATION ou DATE). Les créateurs des tableaux sont vivement encouragés à respecter cette représentation. De ce fait, la plupart des tableaux GFT ont effectivement un nom et les colonnes ont un type, deux indicateurs du contenu sémantique.

Les tableaux de GFT ont principalement trois avantages par rapport aux données tabulaires (tableurs, tables HTML, listes HTML) du web :

1. Leur structure est très simple et claire. Aucune colonne n'est subdi-

visée en sous-colonnes, comme dans les tableurs ;

2. Les colonnes des tableaux GFT sont très souvent typées, ce qui facilite leur annotation sémantique ;
3. Une API simple, efficace et bien documentée permet d'interroger, créer, supprimer et modifier des tableaux en utilisant SQL.

4 Notre approche

Dans cette section, nous détaillons notre approche d'identification des tableaux contenant des données relatives à un concept de l'ontologie et d'extraction des données correspondant à des valeurs d'attributs de ce concept en prenant appui sur un exemple concret d'extraction de données dans GFT portant sur des points d'intérêts (POIs) de la ville (des restaurants et des musées). Comme nous l'avons annoncé en section 1, notre approche comporte trois étapes. Nous décrivons ci-dessous les deux premières étapes. L'étape de peuplement n'est pas décrit. Il s'agit d'un simple processus de traduction des données extraites dans le format de l'ontologie.

4.1 L'identification de tableaux GFT relatives à un domaine donné

Un des challenges importants dans ce travail d'extraction porte sur l'identification des tableaux GFT qui contiennent des données portant sur des concepts spécifiques, dans notre cas sur des restaurants et des musées. Il est évidemment impossible d'effectuer ce travail manuellement vu le grand nombre de tableaux auxquels GFT donne accès. L'algorithme que nous proposons permet d'automatiser cette tâche. Le problème à résoudre est celui-ci : comment peut-on automatiquement déterminer si les données d'un tableau concernent des restaurants ou des musées ? L'approche adoptée est décrite ci-dessous.

GFT contient des millions de tableaux et chaque tableau contient des milliers de lignes. Il est ainsi impossible d'analyser chaque tableau un par un pour sélectionner ceux qui sont pertinents par rapport à nos besoins. Or, il se trouve que le moteur de recherche Google indexe les tableaux de GFT. Ils peuvent alors être recherchés de la même façon que des pages web. Une recherche sur *restaurant* retournera tous les tableaux de GFT contenant le mot *restaurant*. Ceci est très intéressant car cela permet d'obtenir une première sélection de tableaux, même si cette sélection est insuffisante car il ne suffit pas que le mot *restaurant* soit contenu dans un tableau T pour

que toutes les données de T portent effectivement sur des restaurants. Les cas suivants peuvent se produire :

- T contient le mot *restaurant* de façon incidente mais n'a aucune donnée sur des restaurants,
- T contient des données sur des restaurants mais également sur d'autres entités,
- T comprend des données uniquement sur des restaurants.

Quoi qu'il en soit, l'algorithme doit traiter T ligne par ligne pour identifier et retenir uniquement celles qui ont l'information recherchée.

Notre approche de sélection est basée sur l'observation suivante. Il est tout aussi difficile pour un humain de comprendre la sémantique d'un tableau que pour une machine lorsque le contexte n'est pas clair.

Name	Address	Restaurant GUID	Score	URL
A & W	W3047.1A, 3880 - Grant McConachie Way, Richmond BC	db9e966b-97f6-464c-8e3c-314a2b33c10d	0	http://www.foc 8e3c-314a2b3
A & W	FC11 , S Park Royal, West Vancouver BC	1b03ea4e-897a-4e72-8748-9b8b9cf97ce2	2	http://www.foc 4e72-8748-9b
A & W	4, 8671 - No. 1 Road, Richmond BC	1d3e5f8d-9e44-11d6-8952-00a083371b06	1	http://www.foc 8952-00a083
A & W	450 Robson, Vancouver BC	beadb95a-5cd8-4cf1-8824-7dff4ee9aab5	0	http://www.foc 8824-7dff4ee9
A & W - 920 Marine Drive	920 Marine Drive, North Vancouver BC	65f02bb1-13c9-4e58-b1c7-620c95ce1b2c	3	http://www.foc b1c7-620c95c
A & W - Food Court - Level 3	B3842, 3880 - Grant McConachie Way, Richmond BC	bd83ca9f-c917-437e-8233-022d481bc5a7	0	http://www.foc 8233-022d48

FIGURE 2 – Extrait d'un tableau GFT.

Soit, par exemple, l'extrait de tableau GFT de la figure 2. A moins d'être incollable sur Vancouver et le Canada, il n'est pas évident, pour une personne, de comprendre que *A & W - 920 Marine Drive* (5ème ligne, 1ère colonne) est le nom d'un restaurant. Dans ce cas là, quand une personne ne comprend pas à quoi réfère une expression, elle utilise très souvent un moteur de recherche web pour découvrir sa (ses) signification(s) possible(s). Sur cet exemple, l'étude des premiers résultats retournés indique très clairement qu'il s'agit d'un restaurant. Nous avons décidé d'adopter la même

stratégie.

Formellement, le contenu e de chaque cellule $T(i, j)$ d'un tableau est soumis à un moteur de recherche afin d'obtenir une liste de résultats(nous ne considérerons que les 10 premiers [3]). Chaque résultat comporte un *lien* sur une page web, un *titre* et une *courte description*, souvent appelée *snippet* en anglais, qui correspond à un extrait du contenu de la page, généralement celui qui contient les mots clés de la requête. Bien que ces descriptions soient très courtes (30-40 caractères au plus), elles donnent une bonne idée du contenu de la page web. L'analyse du texte de ces descriptions permet alors de déterminer si la page web décrit une certaine entité (restaurant, musée, ...) ou non. Le problème consistant à déterminer si une expression e est une référence à un restaurant est donc un problème de *classification binaire*. Pour chaque description retournée par le moteur de requêtes, nous utilisons un classifieur opérant sur des textes pour déterminer si une description décrit un restaurant. Si la majorité des descriptions trouvées sont considérées comme des descriptions de restaurants, e est considérée comme étant une référence de restaurant et la ième ligne du tableau est sélectionnée. Nous n'affectons pas de mesure de pertinence aux informations classées. Nous donnons plus de détail sur le classifieur en section 5.

4.2 L'extraction de données des tableaux GFT et le peuplement de l'ontologie

Après avoir identifié les lignes du tableau T qui contiennent des données à extraire, dans notre exemple des données sur des restaurants, il faut sélectionner les colonnes qui fournissent des valeurs d'attributs ou de propriétés sur ces restaurants. Nous utilisons, ici, l'ontologie dans laquelle les restaurants sont décrits avec un nom, une adresse, un prix moyen de menu, une catégorie, un numéro de téléphone, une adresse de site web. La recherche de mise en correspondance entre ces attributs et les colonnes des tableaux est basée sur l'application d'heuristiques.

- L'attribut *nom* : La colonne correspondante est trouvée lors de la sélection des lignes contenant des données sur des restaurants (cf. section 4.1).
- L'attribut *adresse* : La colonne correspondante est trouvée en analysant le contenu de chaque colonne dont le type est LOCATION. Étant

3. Nous pouvons obtenir des milliers de réponses. Nos expérimentations nous ont permis de fixer à 10 le nombre de réponses retenues pour avoir une bonne précision.

donnée une expression e dans une cellule d'une colonne ayant le type LOCATION, nous interrogeons le service web *Yahoo ! PlaceFinder* avec e afin de savoir si e correspond à une adresse. *Yahoo ! PlaceFinder* effectue un géocodage de l'adresse en la convertissant en coordonnées géographiques (latitude, longitude).

– Les attributs *prix moyen* et *catégorie*. Les colonnes correspondantes sont déterminées uniquement à partir de leur titre. Autrement dit, une colonne intitulée *prix* ou *coût* contiendra très certainement des valeurs de *prix moyen* alors qu'une colonne intitulée *catégorie*, *type de nourriture* ou *type* est considérée comme contenant des valeurs de *catégorie*.
– Les attributs *numéro de téléphone* et *site web*. Les colonnes correspondantes sont déterminées en utilisant des expressions régulières pour identifier les cellules des tableaux contenant des numéros de téléphone et des URLs.

Mis à part les attributs *prix moyen* et *catégorie*, les colonnes correspondantes sont déterminées après analyse du contenu des colonnes uniquement, le titre des colonnes n'intervient pas. C'est un point important. En effet, toutes les colonnes n'ont pas toujours de titre bien que dans GFT, ce soit relativement rare. Les titres, quand ils existent, sont souvent ambigus et ne reflètent pas bien le contenu sémantique des colonnes. Par exemple, une colonne contenant des noms de restaurants peut être simplement intitulée *Nom* au lieu de *Noms de restaurants*. Les entêtes peuvent, par ailleurs, ne pas être en anglais car GFT est utilisé au niveau mondial. Pour toutes ces raisons, nous nous sommes efforcés d'être le plus possible indépendants des titres des colonnes. Notre approche est encore en cours d'amélioration pour que cette indépendance soit totale.

Une fois les colonnes contenant des valeurs d'attributs de l'ontologie identifiées, les données des lignes ayant été au préalable sélectionnées sont extraites du tableau et insérées dans l'ontologie. Les données obtenues après extraction sont consultables à l'adresse qui suit : `http://sites.google.com/site/gquercini/GFT_data.rdf`. Ces données extraites ont les propriétés suivantes. Il s'agit, pour ce qui concerne les données liées à notre expérimentation, de données portant d'une part sur 1500 restaurants et 500 musées localisés principalement aux États-Unis et en Europe et décrits par leur nom, catégorie, adresse, latitude et longitude Ces données sont issues d'un processus totalement automatique, elles ne sont donc pas exemptes d'erreurs. L'évaluation des erreurs sera partie de notre travail futur.

5 Résultats expérimentaux

Cette section présente les résultats expérimentaux obtenus lors de l'extraction de données de GFT portant sur des restaurants et des musées. Ces résultats concernent l'étape principale décrite en section 4.1, c'est-à-dire l'identification des lignes d'un tableau contenant des données sur des restaurants ou des musées. Cette étape est basée sur l'utilisation d'un classifieur de textes déterminant quels *snippets* obtenus en réponse à une requête posée à un moteur de recherche, décrivent une certaine entité. Notons que *DBpedia* est utilisé pour sélectionner les entités de type restaurant et musée afin de collecter du Web les snippets qui serviront à entraîner le classifieur. Cela ne signifie pas que le classifieur ne peut reconnaître que les entités présentes dans *DBpedia*. Au contraire, il apprend à distinguer les snippets qui portent sur un musée ou un restaurant des snippets qui portent sur d'autres entités.

Le système d'apprentissage utilisé pour la classification est de type *Machines à Vecteurs de Support* (SVM Support Vector Machine) (Cortes & Vapnik (1995)). Plus précisément, nous avons utilisé une Machine à Vecteurs de Support de type C-SVC avec un noyau gaussien. Nous avons fixé deux paramètres : le paramètre C (coût) de la C-SVC et le paramètre γ du noyau. Deux systèmes, notés C_{rest} et C_{mus}, ont été utilisés pour déterminer si un snippet décrit un restaurant ou un musée, respectivement. L'ensemble d'entrainement de C_{rest} (resp. C_{mus}) a été constitué en collectant 300 restaurants (resp. musées) de *DBpedia* ainsi que 300 entités ne correspondant pas à des restaurants (resp. musées). Nous avons utilisé *Microsoft Bing* pour obtenir les *snippets* (10 premiers résultats). Chaque snippet a été converti en un vecteur de valeurs, comme cela doit être fait pour la classification de type Machines à Vecteurs de Support. Pour cela, les *mots vides* (*stopwords*) et les chaînes de caractères contenant des chiffres ou des caractères spéciaux ont été éliminés. Les mots restants ont subi une *racinisation*. Chaque élément du vecteur correspondait donc à un mot et sa valeur à la fréquence relative du mot dans le snippet. Les *snippets* décrivant des restaurants (resp. musées) ont été considérés comme des *exemples positifs* et ceux ne décrivant pas des restaurants (resp. musées) comme des *exemples négatifs*. Les deux systèmes de classification ont été entraînés par validation croisée en 10 blocs. La valeur optimale de C a pu alors être estimée à 32 et celle de γ à 0.5.

Pour l'évaluation, nous avons collecté manuellement 30 grands tableaux de GFT Ces tableaux contiennent 16400 cellules. 608 cellules contiennent des références à des restaurants, 288 cellules sont des références à des mu-

sées. Nous avons mesuré la *précision*, le *rappel* et la *F-mesure* comme suit. Soit G_{rest} (resp. G_{mus}) le nombre de lignes correctement sélectionnées comme étant des données sur des restaurants (resp. des musées) ; la précision est le rapport entre G_{rest} (resp. G_{mus}) et le nombre total de lignes des tableaux ; le rappel est le rapport entre G_{rest} (resp. G_{mus}) et le nombre de lignes portant véritablement sur des restaurants (resp. des musées) ; la F-mesure est la moyenne quadratique combinant la précision et le rappel. Nous avons utilisé yelp.com, un site contenant une base de données très importante sur les restaurants américains pour savoir si une entité correspondait à un restaurant. *DBpedia* a été utilisé pour l'identification des données sur les musées. Les résultats de l'évaluation, effectuée par un seul utilisateur, sont présentés dans le tableau 1.

	Précision	Rappel	F-mesure
Restaurants	0.82	0.74	0.78
Musées	0.68	0.57	0.62

TABLE 1 – Résultats expérimentaux.

La baisse du rappel et de la précision en ce qui concerne les musées est due au fait qu'il est plus difficile de définir des musées que des restaurants. Un musée peut être une galerie mais également une maison historique, un moulin historique, etc. Cette définition est très large et certaines entités du monde réel qui devraient être considérées comme des musées, peuvent être sous-représentées dans l'ensemble d'apprentissage, ce qui explique les différences de résultats entre les deux systèmes de classification C_{rest} et C_{mus}. C_{rest} doit certainement aussi être impacté par cette difficulté à définir un concept mais dans une moindre mesure. Il s'agit d'un problème que nous envisageons d'étudier plus en profondeur.

6 Conclusion

Dans cet article, nous avons présenté une approche d'extraction de données de *Google Fusion Tables* dans l'objectif de peupler une ontologie et de convertir les données obtenues en RDF, de façon à les publier et à faire quelles puissent être exploitées par les applications du web sémantique.

L'aspect le plus complexe de notre approche consiste à déterminer quelles lignes doivent être extraites des tableaux de GFT. Nous avons montré que ceci pouvait être effectué en soumettant le contenu des cellules à un moteur

de recherche. Les résultats de nos premières expérimentations sont encourageants, même s'ils doivent encore être améliorés pour éliminer complètement le bruit.

Références

AMIN M. S., BHATTACHARJEE A. & JAMIL H. (2010). Wikipedia Driven Autonomous Label Assignment in Wrapper Induced Tables with Missing Column Names. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, p. 1656–1660, New York, NY, USA : ACM.

BERNERS-LEE T. (2006). Linked Data - Design Issues. Retrieved July 27, `http://www.w3.org/DesignIssues/LinkedData.html`.

BIZER C., LEHMANN J., KOBILAROV G., AUER S., BECKER C., CYGANIAK R. & HELLMANN S. (2009). DBpedia - A Crystallization Point for the Web of Data. *Web Semant.*, **7**, 154–165.

CAFARELLA M. J., HALEVY A., ZHANG Y., WANG D. Z. & WU E. (2008). WebTables : Exploring the Power of Tables on the Web. *Proc. VLDB Endow.*, **1**(1), 538–549.

COETZEE P., HEATH T. & MOTTA E. (2008). SparqPlug : Generating Linked Data from Legacy HTML, SPARQL and the DOM. In *Linked Data on the Web*.

CORTES C. & VAPNIK V. (1995). Support-Vector Networks. *Mach. Learn.*, **20**, 273–297.

EXPERIAN HITWISE US (2010). Bing - Rising Success (rate). `http://weblogs.hitwise.com/bill-tancer/2010/01/bing_rising_success_rate.html`.

GONZALEZ H., HALEVY A. Y., JENSEN C. S., LANGEN A., MADHAVAN J., SHAPLEY R., SHEN W. & GOLDBERG-KIDON J. (2010). Google Fusion Tables : Web-centered Data Management and Collaboration. In *SIGMOD 2010*, p. 1061–1066, New York, NY, USA : ACM.

GUO X., CHEN Y., CHEN J. & DU X. (2011). ITEM : Extract and Integrate Entities from Tabular Data to RDF knowledge base. In *Proceedings of the 13th Asia-Pacific Web Conference on Web Technologies and Applications*, APWeb'11, p. 400–411, Berlin, Heidelberg : Springer-Verlag.

LE BOURLOUT E. (2012). Knowledge Graph : Google fait un pas vers le Web sémantique. Trouvé en ligne le 18/05/2012.

LIMAYE G., SARAWAGI S. & CHAKRABARTI S. (2010). Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proc. VLDB Endow.*, **3**, 1338–1347.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2007). Yago : a Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, p. 697–706, New York, NY, USA : ACM.

VENETIS P., HALEVY A., MADHAVAN J., PAŞCA M., SHEN W., WU F., MIAO G. & WU C. (2011). Recovering Semantics of Tables on the Web. *Proc. VLDB Endow.*, **4**, 528–538.

ZICARI R. V. (2011). Google Fusion Tables. Interview with Alon Y. Halevy. Retrieved online at `http://www.odbms.org/blog/2011/08/google-fusion-tables-interview-with-alon-y-halevy/`.