Enrichissement contrôlé de bases de connaissances à partir de documents semi-structurés annotés

Yassine Mrabet, Nacéra Bennacer Seghouani, Nathalie Pernelle

Enrichissement contrôlé de bases de connaissances à partir de documents semi-structurés annotés

Yassine Mrabet[1], Nacéra Bennacer[2], Nathalie Pernelle[1]

[1] LRI, UNIVERSITÉ PARIS-SUD, INRIA-SACLAY PCRI, bât. 690, 91405 Orsay
[yassine.mrabet, nathalie.pernelle]@lri.fr

[2] SUPÉLEC, E3S 3 rue Joliot Curie, 91192 GIF-SUR-YVETTE
nacera.bennacer@supelec.fr

Résumé : Grâce au Linked Open Data, les sources RDF mises à disposition sur le Web sont de plus en plus nombreuses. Cependant, ces sources contiennent relativement peu d'information par comparaison au volume d'informations contenues dans les documents semi-structurés. De nombreux outils ont pour objectif d'annoter sémantiquement ces documents mais l'extraction de relations reste une tâche particulièrement difficile quand la structure et le vocabulaire des documents sont hétérogènes. Nous proposons une approche permettant d'enrichir et d'interroger une ou plusieurs bases de connaissances RDF/OWL en exploitant un ensemble de documents sémantiquement annotés. Ces bases sont enrichies par des instances de relations incertaines inférées à partir de la structure des documents, des ontologies et des faits présents dans les bases de connaissances. Une requête SPARQL formulée dans le vocabulaire du domaine est reformulée afin de combiner les faits issus des différentes bases et de trier les réponses en fonction de poids assignés. L'approche a été expérimentée sur des documents HTML et des bases de connaissances issues du Linked Open Data. Les résultats montrent que 63,3% des relations trouvées sont nouvelles avec une précision atteignant 62%.

Mots-clés : Ontologie, Annotation, Enrichissement, Recherche sémantique, Documents semi-structurés, Base de connaissances, RDF/OWL, SPARQL

1 Introduction

L'objectif des projets liés au web de données est de publier de plus en plus de sources de données RDF sur le Web et d'établir des liens entre les données de ces différentes sources. Ces initiatives permettent à des moteurs de recherche sémantiques d'interroger ces sources en combinant éventuellement des données issues de sources différentes et de raisonner

sur ces données. Pourtant, de nombreuses informations restent toujours représentées dans des documents HTML sur le Web. Les outils d'annotation sémantique permettent d'associer à un document ou à des parties de document des métadonnées dont la sémantique est définie dans une ontologie. il est alors également possible de raisonner et d'interroger ces données en se basant sur le vocabulaire de l'ontologie. L'automatisation de ces outils est un facteur clé pour exploiter les informations contenues dans les documents HTML du web.

De nombreux travaux de recherche, issus de domaines complémentaires tels que l'apprentissage, le traitement automatique du langage naturel ou l'ingénierie des connaissances, se sont intéressés à l'extraction d'information ou à l'annotation sémantique automatique de documents. Un grand nombre d'entre eux se sont focalisées sur l'extraction ou l'annotation d'entité nommées ou de termes apparaissant dans les documents (Nadeau & Sekine (2007); Bikel *et al.* (1997)). Ces approches peuvent exploiter des patrons lexico-syntaxiques représentant des expressions régulières apparaissant dans les textes (Suchanek *et al.* (2009); Popov *et al.* (2004); Cimiano *et al.* (2005)) ou des ressources (onto-)lexicales décrivant des ensembles d'entités nommées ou de termes du domaine (Popov *et al.* (2004); Thiam *et al.* (2009)). Certaines de ces approches s'intéressent non seulement à la classification des séquences de mots extraites mais aussi au fait de retrouver l'instance de concept déjà décrite dans une ontologie peuplée (ou base de connaissances) à laquelle l'expression se réfère. Par exemple, l'approche Sofie permet d'identifier une instance de concept en comparant le contexte dans lequel une entité nommée apparaît dans un document (Suchanek *et al.* (2009)) avec les labels des instances de la base de connaissance qui sont liés à l'instance de concept.

Pour exploiter les documents de manière plus riche, il faut également rechercher les instances de relations sémantiques. Certaines approches se basent sur l'existence de patrons lexico-syntaxiques déclarés ou appris sur un corpus de document (Suchanek *et al.* (2009); Aussenac-Gilles & Jacques (2006); Suchanek *et al.* (2006)), d'autres se basent sur l'existence de régularités structurelles ou sur des structures particulières comme les tableaux (Limaye *et al.* (2010); Hignette *et al.* (2009); Buitelaar & Siegel (2006)) ou les infobox de Wikipédia (Suchanek *et al.* (2008); Bizer *et al.* (2009)). Pourtant, quand les documents sont héterérogènes au niveau du vocabulaire et au niveau de leur structure, l'utilisation de patrons ou

de régularités structurelles ne permettra de découvrir qu'un nombre limité de relations. Aussi, il faut pouvoir disposer d'approches complémentaires permettant de découvrir d'autres instances de relations en ne se basant pas sur l'existence de telles régularités.

Dans ce papier, nous présentons l'approche REISA qui permet d'enrichir des bases de connaissances RDF/OWL en utilisant une base de documents HTML annotés par un ou plusieurs outils d'annotations. Pour enrichir les bases de connaissances par des instances de relations incertaines, REISA ne se base pas sur l'existence de patterns structurels ou lexico-syntaxiques mais exploite la proximité des parties de documents référant aux instances de concept (ou représentant des littéraux). Elle exploite également la sémantique de l'ontologie et les faits déjà présents dans la base de connaissance pour contrôler la génération des nouveaux faits. L'incertitude des faits générés par les outils d'annotation automatique ou issus de la méthode d'enrichissement que nous avons définie est représentée en utilisant les graphes nommés RDF. Une requête utilisateur exprimée dans le vocabulaire du domaine est reformulée de façon à atteindre les faits présents dans la base de connaissances initiale, ceux découverts par les outils d'annotations et les faits issus de notre enrichissement. Les réponses sont triées en se basant sur l'incertitude des faits qu'elles contiennent. Cette approche a été evaluée en exploitant deux bases de connaissances issues du Linked Open Data. Ces bases ont été enrichies en utilisant un ensemble de documents HTML décrivant des appels à participation à des conférences scientifiques qui ont été annotés.

Dans la section 2, nous présentons le modèle sémantique qui a été défini pour représenter les bases de connaissances et les bases d'annotations. En section 3, nous présentons l'approche REISA. En section 4, nous présentons les résultats obtenus lors de nos premières expérimentations. Enfin, nous concluons et donnons quelques perspectives.

2 Modèle sémantique d'intégration

Nous proposons un modèle sémantique d'intégration (modèle SIM) permettant de représenter les bases de connaissances et les bases d'annotations (voir figure 1). Il permet de représenter de manière homogène des entités de domaine (i.e. instances de concepts et de propriétés), des parties de documents et les liens entre les entités de domaine et les parties de documents.

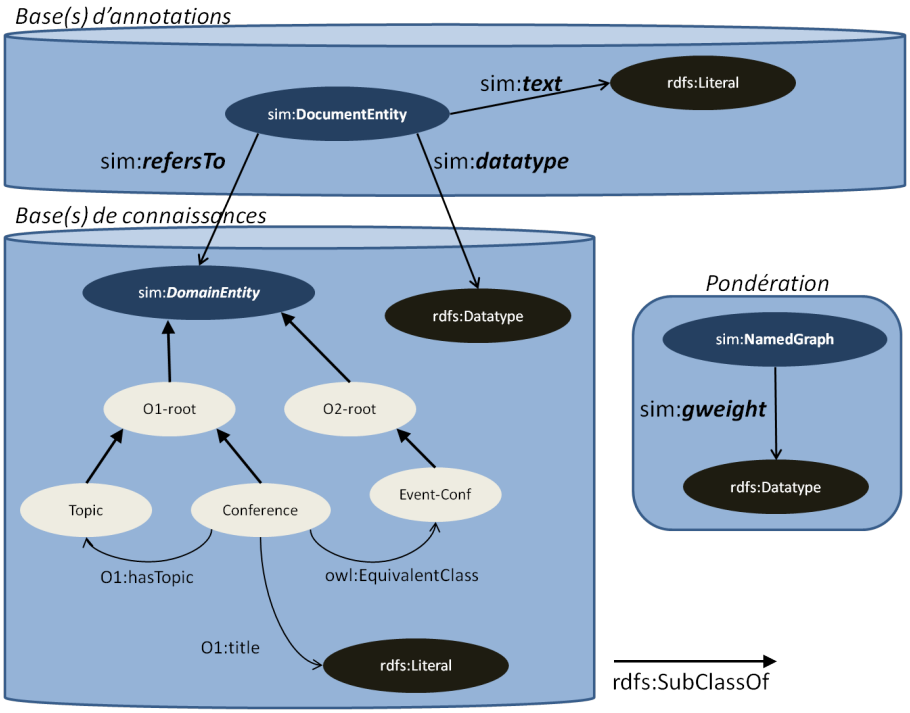

FIGURE 1 – Modèle sémantique d'intégration (SIM).

Base de connaissances. Une Base de Connaissances BC peut être définie par (O, T_O) où O est une ontologie OWL définie par (C_O, R_O, A_O, X_O). C_O est l'ensemble des concepts, R_O est l'ensemble des propriétés définies entre les concepts, A_O est l'ensemble des propriétés définies entre les concepts et les littéraux. X_O est un ensemble d'axiomes déclarant, en particulier, les liens de spécialisation entre concepts et entre propriétés, le domaine et le co-domaine des propriétés, la fonctionnalité ou la fonctionnalité inverse des propriétés et des correspondances sémantiques d'équivalence ou de spécialisation issues de l'alignement avec d'autres ontologies. Une BC contient également l'ensemble T_O des faits décrivant les instances de concepts et de propriétés décrits conformément à O.

Dans le modèle SIM, les concepts de C_O sont des concepts de domaine, un concept de domaine étant représenté par le concept *DomainEntity*. Différentes bases de connaissances $BC_1....BC_n$ peuvent être représentées dans le modèle.

Pondération des bases de connaissances. Nous utilisons les graphes nommés (*NamedGraph*) pour associer une mesure de confiance à un ensemble de faits des BC. Un graphe nommé est un graphe RDF possédant une URI [1]. Le poids d'un graphe nommé représente la confiance qu'un expert, ou le résultat d'une méthode de pondération, a associé aux triplets du graphe nommé. La propriété $sim:gWeight$ permet d'associer à un graphe nommé un poids défini entre 0 et 1. Dans notre approche, un expert attribue un poids de 1 au graphe nommé contenant les faits d'une BC validée et un poids plus faible aux faits d'une BC générée automatiquement par un outil d'extraction ou d'annotation. Les faits générés par notre approche d'enrichissement sont pondérés automatiquement au moment de leur production (c.f. section 3).

Bases d'annotations. Une base d'annotation BA est un ensemble d'entités de documents et de liens trouvés par les annotateurs entre ces entités et les BC. Une entité de document est une instance du concept *DocumentEntity*. Il s'agit d'une partie de document XHTML qui a été sémantiquement annotée. Une entité de document est définie par une URI et un contenu textuel qui lui est lié par la propriété *sim*:*Text*. Il peut s'agir d'un noeud de document défini dans l'arbre DOM du document XHTML (dont le contenu textuel est alors la concaténation des noeuds textuels qui sont fils directs du noeud) ou d'une fenêtre de texte (pouvant correspondre à une entité nommée, un terme ou une séquence de mots quelconque).

Un lien entre une entité de document et une instance de concept de domaine est décrit par la propriété *sim*:*refersTo* : on notera *refersTo(e,i)* le fait qu'une entité de document e réfère l'instance de concept i. Le lien entre une entité de document et un type de base ($rdfs:datatype$) est décrit par la propriété *sim*:*DataType* : *DataType(e,t)* indique que l'entité de document e a pu être typée par t [2]. L'exemple présenté en figure 2 montre un extrait de la base de connaissances de KIM et de la base d'annotations d'un extrait de document.

1. http ://www.w3.org/2004/03/trix/

2. Les types de données définis dans le cadre de XML schema sont des instances de *rdfs*:*dataType*

Extrait de document semi-structuré

 ... **Laos** traces its history to the kingdom of Lan Xang ...
took over **Vientiane** with ... along the Annamite mountains in **Vietnam**.
 ... tools discovered in northern **Laos** attest ...
communities along the **Mekong** River ...   ...
 Following the military defeat of Japan ... the Viet Minh occupied **Hanoi** and proclaimed a provisional government, which asserted national independence ... ...

Extrait de la base de connaissances WKB (KIM).

@prefix graphs: <http://lri.fr/reisa/graphs/>
graphs:knowledgebase = {
kimkb:Laos.0 **rdf**:**type** onto:Country
kimkb:Laos.0 **onto**:**partOf** kimkb:Continent.2
kimkb:Continent.2 **rdf**:**type** onto:Continent
kimkb:Continent.2 **rdfs**:**label** "Asia"
kimkb:Vientiane.0 **rdf**:**type** onto:City
kimkb:Vientiane.0 **onto**:**capital** kimkb:Laos.0 }

Base d'annotations de l'extrait de document.

graphs:annotationsbase = {
corpus:doc0/html/body/div/p[3]/a.0 **rdf**:**type** sim :DocumentEntity
corpus:doc0/html/body/div/p[3]/a.0 **sim**:**refersTo** kimkb:Vietnam.0
corpus:doc0/html/body/div/p[3]/a.0 **sim**:**text** "Vietnam"
corpus:doc0/html/body/div/p[3].12 **sim**:**refersTo** kimkb:Laos.0
corpus:doc0/html/body/div/p[3].12 **sim**:**text** "Laos"
corpus:doc0/html/body/div/p[3].20 **sim**:**refersTo** kimkb:Mekong.0
corpus:doc0/html/body/div/p[3].20 **sim**:**text** "Mékong"
corpus:doc0/html/body/div/p[3]/a[2].0 **sim**:**refersTo** kimkb:Hanoi.0
corpus:doc0/html/body/div/p[3]/a[2].0 **sim**:**text** "Hanoi"
}
graphs:knowledgebase sim:gweight 1
graphs:annotationsbase sim:gweight 0.9

FIGURE 2 – Extrait d'un document semi-structuré, de ses annotations et de la base de connaissances WKB.

3 Approche REISA

L'approche REISA (contRoled Enrichment and Interrogation of Semantic Annotations) est constituée de trois principaux modules (cf. figure 3) : (1) un module d'intégration qui permet de représenter les BC et les BA conformément au modèle SIM, (2) un module d'enrichissement qui exploite les BC, la BA, et la structure des documents pour génerer de nouvelles instances de propriétés de domaine, (3) un module d'interrogation et de reformulation qui exploite la base d'enrichissement et les BC pour répondre aux requêtes des utilisateurs.

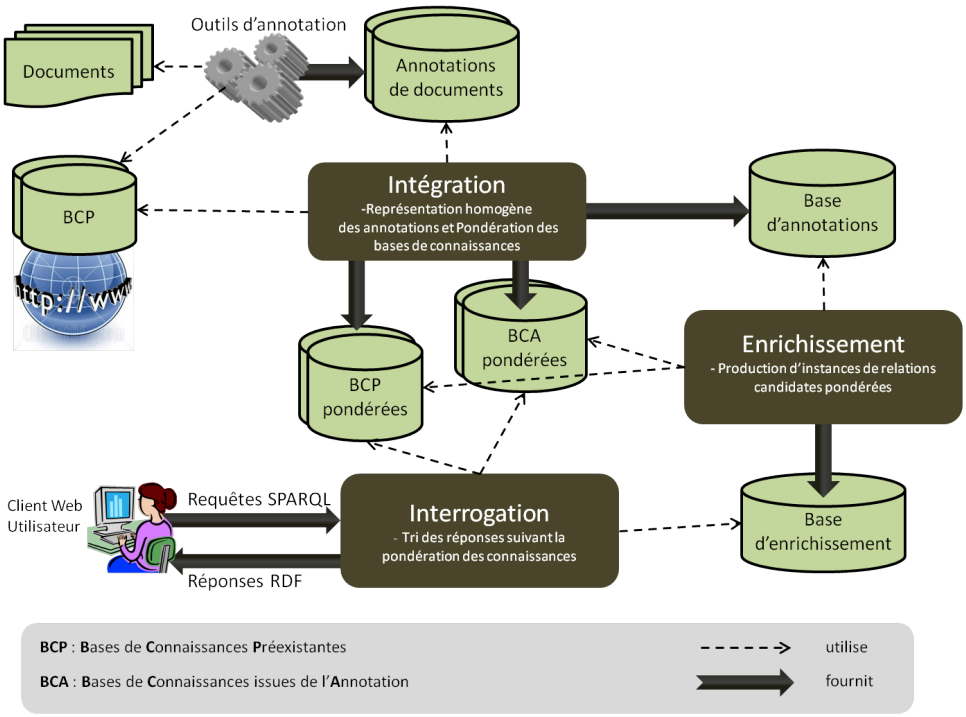

FIGURE 3 – Architecture de REISA.

3.1 Construction de la base d'enrichissement

Pour extraire de nouvelles instances de propriétés, nous exploitons les hypothèses suivantes : (1) plus les informations sont proches dans les documents semi-structurés, plus elles sont susceptibles d'être liées sémantiquement et (2) la structure des documents peut rendre compte d'une partie de leur sémantique. Cependant, se fonder uniquement sur ces hypothèses risque de générer de nombreux candidats erronés.

Dans notre approche, nous exploitons la proximité structurelle entre les entités de document pour proposer de nouvelles propriétés de domaine entre les instances de concepts (ou entre instances de concepts et littéraux) qui sont référencées par ces entités. Nous pondérons ces propriétés candidates par une mesure de confiance qui est calculée en fonction de la distance entre les entités de documents annotées. De plus, nous exploitons la sémantique des ontologies et les instances de propriétés déjà présentes dans les BC pour contrôler au mieux les faits candidats. Dans cette sec-

tion, nous présentons tout d'abord, la notion de voisinage sémantique, puis la méthode de construction de la base d'enrichissement.

Voisinage sémantique. Deux instances de concept i_1 et i_2 sont dites sémantiquement voisines à distance d pour une propriété P, noté $V_P^d(i_1, i_2)$, lorsqu'il existe deux entités de documents e_1 et e_2 qui référent à ces instances, que ces entités de document sont à distance d, et que les types des instances sont domaine et co-domaine de P :
$refersTo(e_1, i_1) \wedge refersTo(e_2, i_2) \wedge (distance(e_1, e_2) = d) \wedge$
$type(i_1, C_1) \wedge type(i_2, C_2) \wedge domain(P, C_1) \wedge range(P, C_2) \rightarrow V_P^d(i_1, i_2)$
De même, une instance de concept et le texte d'une entité de document sont dits sémantiquement voisins à distance d pour une propriété P si le type de littéral associé à ce texte et le concept de l'instance sont respectivement co-domaine et domaine de P.
$refersTo(e_1, i_1) \wedge datatype(e_2, T) \wedge (distance(e_1, e_2) = d) \wedge type(i_1, C_1)$
$\wedge domain(P, C_1) \wedge range(P, T) \wedge text(e_2, l_2) \rightarrow V_P^d(i_1, l_2)$
La distance entre deux entités de document correspond à la distance qui les sépare dans l'arbre DOM du document (i.e. le plus court chemin).

La base d'enrichissement est construite en utilisant les bases d'annotations, les ontologies et les bases de connaissances. Nous notons G^p le graphe nommé de poids p. Les graphes nommés de poids 1 sont considérés comme une référence qui regroupe l'ensemble des faits provenant des bases de connaissances sûres.

Graphe nommé de voisinage. Un graphe nommé de voisinage, noté $G^{w(d)}$, regroupe les faits produits par enrichissement à partir d'entités de document distantes de d dans le document. Nous définissons le poids d'un graphe nommé de voisinage, $w(d)$, de manière à ce qu'il soit inversement proportionnel à la distance d. Il est calculé de la façon suivante pour α une constante strictement inférieure à 1 et fixée par l'expert :

$$w(d) = \frac{\alpha}{d+1} \tag{1}$$

Une instance de propriété $P(i_1,i_2)$, inférée entre une instance de concept i_1 et une instance de concept i_2, est ajoutée au graphe $G^{w(d)}$ ssi :

- V_P^d (i_1,i_2)
  i_1 et i_2 sont sémantiquement voisins à distance d

- $\neg\exists$ P$(i_1,i_2) \in G^p$ tq. $p > w(d)$

Le fait n'existe pas dans les graphes de meilleur poids.

– Si P est fonctionnelle alors $\neg\exists\ z$ tq. $P(i_1, z) \in G^1$
La propriété étant fonctionnelle, si i_1 est déjà lié à une instance de concept z dans la base de connaissances sûre, soit z est une référence au même objet du monde réel et il ne s'agit donc pas d'une nouvelle connaissance, soit il s'agit d'un fait candidat erroné.

– Si P est inverse fonctionnelle alors $\neg\exists\ z$ tq. $P(z, i_2) \in G^1$

Une instance de propriété $P(i,l)$, inférée entre une instance de concept i et un littéral l, est ajoutée au graphe $G^{w(d)}$ ssi :

– V_P^d (i,l)
– $\neg\exists$ P$(i,l) \in G^p$ tq. $p > w(d)$
– Si P est fonctionnelle alors $\neg\exists\ m$ tq. P$(i,m) \in G^1$

Une instance de propriété $P(x,y)$, inférée entre une instance de concept x et une instance ou un littéral y, est ajoutée au graphe $G^{w(d)}$ si :

– $\neg\exists$ P$(x,y) \in G^p$ tq. $p > w(d)$
– Si P est fonctionnelle alors $\neg\exists\ z$ tq. $P(x, z) \in G^1$
– Si P est inverse fonctionnelle alors $\neg\exists\ z$ tq. $P(z, y) \in G^1$
– $\exists\ P'$ tq. $subPropertyOf(P', P)$ et $P'(x,y) \in G^{w(d)}$
Si une sous-propriété de P lie x et y dans $G^{w(d)}$, $P(x, y)$ est ajouté à $G^{w(d)}$. Les instances de propriétés sont propagées avec le même poids. La probabilité d'appartenance à un ensemble est au moins aussi grande que la probabilité d'appartenance à un sous-ensemble (Zadeh (1965)).

Nous construisons les faits incertains et les graphes nommés de voisinage associés jusqu'à un seuil de distance μ fixé. L'algorithme d'enrichissement construit progressivement les faits candidats en exploitant les distances de voisinage les plus courtes d'abord.

L'exemple décrit dans la figure 4 montre les graphes de voisinage résultant du module d'enrichissement appliqué à la base d'annotations et à l'extrait de la base de connaissances WKB de l'exemple précédent (cf. figure 2). Les poids ont été calculés avec α=0.9 et un seuil de distance μ=3 (distances structurelles 0, 1 et 2).

REISA a pu générer trois instances de propriétés candidates. Par exemple, la propriété *kim*:*capital* a pu être inférée entre "Hanoi" et "Vietnam" qui étaient à distance 2 dans le document mais non entre "Vientiane" et "Vietnam" qui étaient à distance 0. Cela est dû à la fonctionnalité de la propriété *kim*:*capital* qui nous a permis d'éviter de proposer la ville "Vientiane"

```
@prefix graphs: <http://lri.fr/reisa/graphs/>
graphs:candidates.distance.0 {
kimkb:Mekong.0 dbpedia:country kimkb:Laos.0
}
graphs:candidates.distance.1 {
kimkb:Mekong.0 dbpedia:country kimkb:Vietnam.0
}
graphs:candidates.distance.2 {
kimkb:Hanoi.0 kim:capital kimkb:Vietnam.0
}
graphs:candidates.dbpedia.country.0 sim:gweight 0.9
graphs:candidates.dbpedia.country.1 sim:gweight 0.45
graphs:candidates.kim.capital.2 sim:gweight 0.3
```

FIGURE 4 – Exemple de trois graphes de voisinage issus de l'enrichissement.

comme capitale du "Vietnam", puisque la ville est déjà connue comme capitale d'un autre pays, le Laos, dans la BC (cf. figure 2).

3.2 Reformulation des requêtes utilisateur

Les requêtes utilisateur, formulées avec le vocabulaire des ontologies de domaine, sont ré-écrites pour interroger les bases de connaissances et la base d'enrichissement tout en triant les réponses en fonction des poids des graphes nommés auxquels appartiennent les faits. Pour une requête Sparql donnée, la reformulation consiste à :

- Récupérer le nom du graphe nommé auquel appartient chaque patron de triplet et le poids de ce graphe en utilisant la propriété *sim*:*gweight*.
- Trier les réponses avec une fonction d'ordre appliquée à l'ensemble des poids.

La figure 5 présente un exemple de cette reformulation pour une requête utilisateur demandant la liste des pays avec leur capitale. Dans cette réécriture, la fonction moyenne est utilisée comme exemple pour trier les réponses suivant les poids des graphes nommés dont elles sont issues mais d'autres fondtions d'agirégation peuvent être utilisées.

4 Éxpérimentation et évaluation

Nous présentons dans cette section les premières expérimentations que nous avons menées pour évaluer l'approche REISA. L'objectif de ces ex-

SELECT ?c ?t WHERE {
?c rdf :type kim:Country
?t rdf :type kim:City
?t kim:capital ?c
}
	SELECT ?c ?t WHERE {
	GRAPH ?g1 { ?c rdf :type kim:Country }
	GRAPH ?g2 { ?c rdf :type kim:City }
	GRAPH ?g3 { ?t kim:capital ?c }
	?g1 sim:gweight ?p1
	?g2 sim:gweight ?p2
	?g3 sim:gweight ?p3
	} ORDER BY avg(?p1, ?p2, ?p3)

FIGURE 5 – Exemple de requête reformulée.

périmentations est d'évaluer la précision des faits incertains générés par notre méthode d'enrichissement en fonction de (i) leurs poids (ou distance de voisinage) et (ii) l'utilisation ou non des contraintes exploitant les faits et les axiomes présents dans les bases de connaissances sûres.

Bases de connaissances. Nous avons formé une base de connaissances sûre BC_1 en combinant des extraits de la base de connaissances DBLP-RDF [3] qui décrit des références bibliographiques et des extraits de la base de connaissances WKB de KIM [4] qui est généraliste. DBLP-RDF utilise les ontologies *SWRC* et *FOAF* comme référence. *WKB* utilise l'ontologie *PROTON* comme référence. Dans nos expérimentations nous avons utilisé l'union de ces deux ontologies [5].

Plus précisément, l'ontologie de BC_1 représente les articles avec le concept *swrc*:*InProceedings*, les auteurs avec le concept *foaf*:*Agent* et son équivalent *kim*:*Person*, les actes avec le concept *swrc*:*Proceedings*), les séries avec le concept *swrc*:*Conference*, les conférences avec le concept *reisa*:*Event* et pour chaque conférence, le pays avec le concept *kim*:*Country*, la ville avec le concept *kim*:*City* et de manière générale son lieu avec le concept *kim*:*Location*.

Pour représenter le lien entre les conférences et leurs lieux, nous avons défini la propriété *reisa*:*hasLocation* et ses sous-propriétés fonctionnelles *reisa*:*hasCountry* et *reisa*:*hasCity*.

Les faits de la base de connaissance BC_1 sont les suivants :

– Actes des événements qui se sont déroulés entre 2003 et 2007 de DBLP RDF (5608 instances de *swrc*:*Proceedings* et leurs descrip-

3. http ://thedatahub.org/dataset/fu-berlin-dblp
4. http ://www.ontotext.com/kim/semantic-annotation
5. http ://www.lri.fr/∼mrabet/reisa-onto.rdf

tions
- Villes, pays et localisations décrits dans KIM *WKB* et leurs descriptions (12484 instances de *kim*:*Location*, 3093 de *kim*:*City*, 502 de *kim*:*Country*) .
- Instances d'événements construits à partir des instances de *swrc*: *Proceedings* et décrits par une date (la date du proceeding), un titre (titre de la série concernée) et éventuellement une ville et un pays, quand l'information est présente dans le titre DBLP de la conférence (5608 instances de *reisa*:*Event*).

Base d'annotations. Nous avons extrait 511 documents HTML du web décrivant 57 sites d'appels à communication en langue anglaise [6]. Nous nous sommes focalisés sur la recherche des lieux où se déroulent les conférences sachant que cette information n'est pas toujours renseignée dans BC_1. Nous avons annoté les pays et les villes avec l'outil d'annotation de la plateforme KIM. Les événements et leurs dates ont été annotés en utilisant des patterns lexico-syntaxiques que nous avons définis. L'application de ces outils d'annotation a permis de découvrir 1429 entités de documents référant à des événements, 840 entités référant à des villes et 1618 entités référant à des pays. Les outils d'annotation utilisés n'ont fourni aucune instance de propriété. Sur les 1429 entités de document annotées, 348 réfèrent à des instances de BC_1 et 1081 entités réfèrent à de nouvelles instances découvertes dans les documents qui sont réunies dans une base de connaissances BC_2 dont le poids a été fixé à 0.9. Cette valeur de poids indique que les faits de ce graphe sont jugés moins sûrs que les faits de BC_1.

Enrichissement. Nous avons appliqué l'appoche REISA avec les bases de connaissances BC_1 et BC_2 et le corpus annoté de 511 documents, ce qui a permis de retrouver 187 propriétés (instances des propriétés *reisa*:*hasCountry*, *reisa*:*hasCity* et *reisa*:*hasLocation*), avec une limite de distance de voisinage μ fixée à 2 et α fixé à 0.9.

Résultats. Nous donnons la précision des faits de la base d'enrichissement pour les propriétés fonctionnelles *reisa*:*hasCountry*, *reisa*:*hasCity* et *reisa*:*hasLocation* (i.e. le nombre de faits corrects divisé par le nombre de faits trouvés). La justesse des différents lieux trouvés pour les conférences a été vérifiée en examinant les pages HTML les décrivant.

Nous avons défini différents tests pour évaluer la précision des faits incertains en fonction des critères utilisés pour la construction des graphes

6. http ://www.lri.fr/m̃rabet/reisa-dataset.zip

V	Voisinage sémantique uniquement
VF	V + critère de fonctionnalité
VR	V + élimination des propriétés existantes dans BC_1
VRF (REISA)	VR + critère de fonctionnalité

TABLE 1 – Tests définis pour l'évaluation.

	hasCity			hasCountry		
	d=0	d=1	d=2	d=0	d=1	d=2
V	45,5 (10)	38,2 (13)	28,1(16)	72,0(18)	44,0(22)	31,4(32)
VR	36,8 (7)	35,7 (10)	21,7 (10)	58,8 (10)	33,3 (14)	22,2 (20)
VF	46,6 (7)	34,5 (10)	23,8 (10)	78,6 (11)	83,3 (15)	41,5 (22)
REISA	46,7 (7)	34,5 (10)	23,8 (10)	76,5 (10)	77,8 (14)	39,2 (20)
	hasLocation					
	d=0		d=1		d=2	
V	60,4 (29)		42,8 (36)		30,8 (49)	
VR	47,4 (18)		35,2 (25)		22,6 (31)	
REISA	47,4 (18)		35,2 (25)		22,6 (31)	

TABLE 2 – Précision (nb de faits corrects) par propriété, distance et test.

nommés issus de l'enrichissement (cf. tableau 1). Le tableau 2 présente la précision (en %) et le nombre de faits corrects trouvés (par propriété, par distance et par test).

La qualité (précision) de ces nouveaux faits varie en fonction de la distance de voisinage. A distance 0 (même nœud de document), pratiquement un fait sur 2 est correct pour les villes et 76.5% des faits sont corrects pour les pays. La précision diminue notablement quand la distance augmente. Cependant, nous notons que pour la relation *hasCountry* la précision augmente à distance 1 par rapport à la distance 0 (elle passe de 76,5% à 77,8%). Cela est dû à deux éléments : (i) de nouvelles réponses correctes sont retrouvées à distance 1 et (ii) le critère de fonctionnalité élimine plus de mauvaises réponses à distance 1 qu'à distance 0. Les résultats montrent également que le critère de fonctionnalité a un impact important sur la précision des réponses. Ainsi, pour la propriété $hasLocation$, ce critère améliore la précision de 47,4% à 62,1% à distance 0. Le tableau 3 présente le pourcentage de faits nouveaux (non existants dans BC_1) générés par notre approche par rapport au nombre total de faits trouvés (i.e. faits de la base d'enrichissement. Le nombre de réponses nouvelles obtenues grâce à notre approche d'enrichissement varie entre 55 et 70 % des faits $hasCity$, $hasCountry$ et $hasLocation$ générés par enrichissement à distance 0.

Les résultats montrent l'intérêt de l'exploitation des bases de connais-

hasCity			hasCountry			hasLocation		
d=0	d=1	d=2	d=0	d=1	d=2	d=0	d=1	d=2
70,0%	76,9%	62,5%	55,6%	63,6%	62,5%	62,1%	69,4%	63,3%

TABLE 3 – % de faits nouveaux parmi les faits retrouvés (VRF/V).

sances pour le contrôle de l'enrichissement. Les effets sur la précision des faits construits sont visibles même si tous les liens de référence n'ont pas pu être établis entre les entités de document annotées et les instances de concept de BC_1. En effet, plusieurs entités de document ont des instances de concept correspondantes dans la base de connaissances sûre BC_1 mais les outils d'annotation utilisés n'ont pas permis d'établir ce lien. Dans ce travail, notre objectif n'est pas d'évaluer l'efficacité des outils d'annotation et de désambiguïsation. Cependant, le contrôle que nous effectuons sera plus performant si plus de liens de référence sont découverts et désignent des instances de concept décrites dans des bases de connaissances sûres.

5 Conclusion

Dans cet article, nous avons présenté l'approche REISA pour l'enrichissement contrôlé de bases de connaissances à partir de documents semi-structurés portant sur un domaine cible. Cette méthode d'enrichissement est complémentaire aux méthodes d'extraction «classiques» des relations sémantiques qui utilisent, par exemple, des patrons lexico-syntaxiques ou des classifieurs automatiques. Elle génère des faits incertains à partir des documents semi-structurés et des bases de connaissances disponibles sans requérir une quelconque régularité lexico-syntaxique ou structurelle dans les documents. Notre approche tient compte de la qualité des bases de connaissances exploitées. Cette qualité est estimée suivant la confiance accordée à la source des connaissances ou suivant leur méthode de production. Nous distinguons, en particulier, les bases de connaissances sûres du Linked Open Data, les bases de connaissances produites par l'annotation automatique du corpus de domaine ciblé et la base d'enrichissement construite par REISA.

Pour représenter l'incertitude des faits, nous avons utilisé les graphes nommés RDF qui permettent d'associer des palliers de valeurs de confiance (poids) à un ensemble de faits et d'optimiser la construction de la base d'enrichissement et l'interrogation par comparaison à une représentation en RDF flou (e.g. Straccia (2009)).

Notre approche peut être exploitée pour (i) le peuplement (semi-) automatique de bases de connaissances en faisant valider les faits incertains par des experts du domaine (ii) la recherche sémantique d'informations sur un corpus cible pertinent pour un domaine d'intérêt (e.g médecine, biologie, cinéma) et (iii) l'intégration de sources différentes grâce aux nouvelles relations de domaine trouvées, reliant des instances appartenant à des bases de connaissances différentes.

À court terme, nous comptons expérimenter l'approche REISA sur plus de documents et sur un autre domaine. Nous souhaitons également permettre à un expert de définir des règles spécifiques au domaine autres que la fonctionnalité (inverse) des propriétés afin de contrôler la cohérence des faits générés quand ses règles sont appliquées (ex : la ville d'une conférence est située dans le pays où se déroule cette conférence).

Références

AUSSENAC-GILLES N. & JACQUES M.-P. (2006). Designing and evaluating patterns for ontology enrichment from texts. In *International conference on Knowledge Engineering and Knowledge Management (EKAW)*, p. 158–165.

BIKEL D. M., MILLER S., SCHWARTZ R. & WEISCHEDEL R. (1997). Nymble. In *Proceedings of the fifth conference on Applied natural language processing*, p. 194–201, Morristown, NJ, USA : Association for Computational Linguistics.

BIZER C., LEHMANN J., KOBILAROV G., AUER S., BECKER C., CYGANIAK R. & HELLMANN S. (2009). Dbpedia – a crystallization point for the web of data. *journal of Web Semantics : Science, Services and Agents on the World Wide Web*, **7**, 154–165.

BUITELAAR P. & SIEGEL M. (2006). Ontology-based information extraction with soba. In *International Conference on Language Resources and Evaluation (LREC)*, p. 2321–2324.

CIMIANO P., LADWIG G. & STAAB S. (2005). Gimme' the context : context-driven automatic semantic annotation with c-pankow. In *WWW conference*, p. 332–341.

HIGNETTE G., BUCHE P., DIBIE-BARTHÉLEMY J. & HAEMMERLÉ O. (2009). Fuzzy annotation of web data tables driven by a domain ontology. In *Extended Semantic Web Conference (ESWC)*, p. 638–653.

LIMAYE G., SARAWAGI S. & CHAKRABARTI S. (2010). Annotating and searching web tables using entities, types and relationships. *VLDB journal*, **3**, 1338–1347.

NADEAU D. & SEKINE S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, **30**(1), 3–26. Publisher : John Benjamins Publishing Company.

POPOV B., KIRYAKOV A., KIRILOV A., MANOV D., OGNYANOFF D. & GORANOV M. (2004). Kim - semantic annotation platform. *Journal of Natural Language Engineering*, **10**(3), 375–392.

STRACCIA U. (2009). A minimal deductive system for general fuzzy rdf. In *International Conference on Web Reasoning and Rule systems (RR)*, p. 166–181.

SUCHANEK F., SOZIO M. & WEIKUM G. (2009). Sofie : A self-organizing framework for information extraction. In *WWW conference*, p. 631– 640.

SUCHANEK F. M., IFRIM G. & WEIKUM G. (2006). Combining linguistic and statistical analysis to extract relations from web documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, p. 712, New York, USA : ACM Press.

SUCHANEK F. M., KASNECI G. & WEIKUM G. (2008). Yago : A large ontology from wikipedia and wordnet. *Journal of Web Semantics*, **6**(3), 203–217.

THIAM M., BENNACER N., PERNELLE N. & LO M. (2009). Incremental ontology-based extraction and alignment in semi-structured documents. In *DEXA*, p. 611–618.

ZADEH L. (1965). Fuzzy sets. *Information and Control*, p. 338–353.