Apport du Web et du Web de Données pour la recherche d'attributs

Rafik Abbes — Arlind Kopliku — Karen Pinel-Sauvagnat — Nathalie Hernandez — Mohand Boughanem

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG
118 route de Narbonne, F-31062 Toulouse Cedex 9, France

rafik.abbes@irit.fr, akopliku@weboglobin.com, karen.Sauvagnat@irit.fr, nathalie.hernandez@irit.fr, mohand.boughanem@irit.fr

RÉSUMÉ. *Nous nous intéressons dans cet article aux requêtes de type entité pour lesquelles on souhaite renvoyer un ensemble d'attributs (propriétés) et leurs valeurs. Ces attributs peuvent être collectés à partir de plusieurs sources et agrégés dans un seul document. Par exemple l'entité "France" peut avoir les attributs "Langue officielle: Français", "Villes:Paris, Toulouse, Lyon, ..." et "Population:* 65350000 *(en 2012)". Un attribut peut être monovalué ou multivalué, et peut éventuellement dépendre d'autres dimensions. Pour chercher les attributs d'une entité, nous avons exploité deux sources: les tables relationnelles du Web (issues du HTML) et le Web de Données. Afin d'évaluer le potentiel de ces sources, nous avons mis en place une évaluation utilisateur. Les analyses ont montré l'utilité de combiner ces deux sources pour répondre aux requêtes de type entité.*

ABSTRACT. *In this paper, we aim at answering entity-queries by searching their relevant attributes that can be collected from several sources and aggregated into a single document. For example, the entity "France" can have attributes such as "official language: French", "Cities:Paris, Toulouse, Lyon, ..." and "Population:* 65350000 *(in 2012)". An attribute can be multivalued or monovalued, and may possibly depend on other dimensions. To search entity attributes, we used two sources: relational web tables and Linked Data. To assess the contribution of these sources, we made a user study. Analyzes showed the usefulness of combining these two sources to answer entity-queries.*

MOTS-CLÉS : *recherche d'attributs, requête de type entité, tables relationnelles du Web, Web de Données*

KEYWORDS: *attribute retrieval, entity-queries, relational web tables, Linked Data*

article soumis à *CORIA et RJCRI*

1. Introduction

Selon une étude récente (Pound *et al.*, 2010), plus de la moitié des requêtes du Web ciblent une entité particulière ou des entités d'une classe. Une entité est une "chose" qui peut être distinctement identifiée (Chen, 1976) : *"président Barack Obama"*, *"IPhone 5"*, *"musée du Louvre à Paris"*. Une classe d'entités est un ensemble composé d'entités de même type : *"Président"*, *"Smartphone"*, *"Musées de France"*. Pour répondre à ces requêtes, une alternative à la liste traditionnelle de documents consiste à agréger dans un seul document toutes les informations retrouvées sur l'entité ou la classe d'entités (des descriptions, des images, des vidéos, des caractéristiques sous forme d'attributs, des avis, etc) issues de plusieurs sources (documents, bases de connaissances, base de données ...).

Nous nous intéressons particulièrement dans cet article aux requêtes de type entité pour lesquelles nous cherchons à décrire une entité en renvoyant tous ses attributs pertinents. Un attribut est un couple nom-valeur(s) permettant de décrire une entité. L'entité *"Barack Obama"* peut avoir par exemple l'attribut *"Date de naissance : 4 août 1961"*. Un exemple de réponse pour la requête *"IPhone 5"* est illustré dans la figure 1. Les attributs résultats sont présentés dans un tableau. Nous voyons qu'un attribut peut être monovalué (*"Fabricant : Apple"*) ou multivalué (*"Réseaux : 2G, 3G, 3G+, 4G"*) et peut éventuellement dépendre d'une autre dimension comme le temps (*"Nombre d'exemplaires vendus : 5 millions (Septembre 2012)"*). Cette variété de types d'attributs rend la tâche de recherche d'attributs d'une entité plus complexe.

IPhone 5	
Fabricant	Apple
Réseaux	2G, 3G, 3G+, 4G
Nombre d'exemplaires vendus	5 millions (Septembre 2012)
Site internet	www.apple.com
Date de sortie	21-09-2012
Poids	112 gr
...	

Figure 1. *Exemple de réponse pour la requête IPhone 5*

Le Web représente la source la plus utilisée pour la recherche d'attributs. Les attributs d'une entité peuvent se trouver dans le Web sous forme de contenu non structuré (texte brut) ou structuré (tables HTML, listes, . . .). Par exemple, plusieurs méthodes ont été proposées afin d'extraire les attributs pertinents d'une entité donnée en exploitant les documents Web (Kopliku *et al.*, 2011, Cafarella *et al.*, 2008).

Une autre source pouvant fournir des attributs est le Web de Données. A la différence du Web classique qui utilise des documents, le Web de données manipule des ressources identifiées par des URIs[1](Bizer *et al.*, 2008). Il consiste en un ensemble de données structurées conçues pour être exploitées non seulement par

1. Uniform Resource Identifier

des humains mais aussi par des machines. Ces données sont structurées sous forme de triplets RDF[2] comportant chacun un sujet, un prédicat et un objet. Elles peuvent être extraites via des programmes informatiques en utilisant le langage SPARQL[3]. Aujourd'hui, le volume de données disponibles ne cesse d'augmenter (Bizer *et al.*, 2009). Pour mettre en pratique les concepts du Web de Données, le W3C[4] a lancé le projet *Linking Open Data*[5]. Ce projet contient 295 entrepôts de triplets RDF (datasets) comme DBpedia[6], Freebase[7], LinkedMDB[8], etc. La figure 2 montre un extrait du nuage de datasets du projet *Linking Open Data*. Un dataset peut utiliser ou référencer une ressource d'un autre dataset ce qui permet de lier les datasets entre eux.

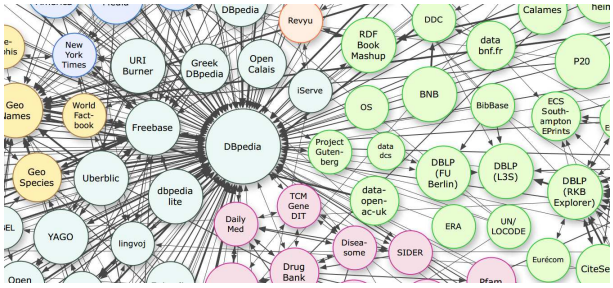

Figure 2. *Extrait du nuage du projet Linking Open Data, Septembre 2011*

A notre connaissance, le Web de données n'a que peu été utilisé pour la recherche d'attributs correspondant à une entité donnée. Dans le but d'estimer son intérêt par rapport au Web classique (et plus particulièrement aux tables HTML du Web classique), nous avons mis en place une évaluation utilisateur. Nous cherchons plus précisément à répondre aux questions suivantes :

– le Web de Données permet-il d'ajouter de l'information pertinente pour répondre aux besoins utilisateur ? Est-t-il important de combiner les données du Web de Données avec celles du Web classique ?

– Si cette combinaison est utile, à quelles éventuelles problématiques devrons-nous faire face si nous devions construire une système de recherche d'information prenant en entrée des requêtes entité ?

Cet article est organisé comme suit. Nous commençons par présenter dans la section 2 un état de l'art sur la recherche d'attributs. La section 3 présente les sources et les données utilisées dans le cadre de notre évaluation. La section 4 décrit la procédure

2. Resource Description Framework
3. SPARQL Protocol and RDF Query Language
4. World Wide Web Consortium
5. http ://richard.cyganiak.de/2007/10/lod/
6. http ://dbpedia.org/
7. http ://www.freebase.com/
8. http ://www.linkedmdb.org/

d'évaluation et les requêtes choisies. Nous analysons dans la section 5 les résultats obtenus et nous concluons et énonçons quelques perspectives en section 6.

2. État de l'art : Recherche d'attributs

Le Web est la source la plus utilisée pour l'extraction des attributs. La méthode proposée dans (Bellare *et al.,* 2007) permet d'extraire les attributs d'une entité à partir des articles Web en utilisant des patterns lexico-syntaxiques comme le pattern "the x of y" qui peut être utilisé pour identifier les attributs d'une entité comme par exemple dans la phrase "the capital of France is Paris". Dans (Pasca *et al.,* 2007), les auteurs extraient les attributs d'une classe d'entités en appliquant des patterns lexico-syntaxiques sur des logs de moteurs de recherche. Dans (Tokunaga *et al.,* 2005), la méthode proposée permet d'acquérir des attributs d'une classe d'entités à partir des documents Web en exploitant des patterns lexico-syntaxiques, des statistiques sur les termes, et les balises HTML.

D'autres méthodes d'extraction d'attributs (Kopliku *et al.,* 2011, Cafarella *et al.,* 2008, Chen *et al.,* 2000) exploitent les tables relationnelles du Web. Ces tables sont définies comme étant des tables HTML contenant des données relationnelles. La figure 3 en montre un exemple. Par opposition, les tables utilisées à des fins de mise en page ne sont pas relationnelles. Les auteurs utilisent des classifieurs permettant de détecter les tables relationnelles en utilisant l'apprentissage supervisé, puis exploitent ces tables pour extraire les entités et leurs attributs.

A notre connaissance, peu d'approches ont utilisé le Web de Données pour extraire des attributs dans un but de recherche d'information. On peut citer (Krichen *et al.,* 2011), dans laquelle une approche a été proposée pour la recherche d'attributs pertinents d'une entité ou une classe en se basant sur la base de connaissance DBpedia. Les attributs sont extraits puis triés en fonction du nombre d'entités d'une même classe pour laquelle ils étaient renvoyés. Cependant, la couverture de cette approche reste limitée puisqu'elle ne s'appuie que sur une seule source.

La recherche d'attributs n'est pas utile qu'en RI. Par exemple, faire des statistiques sur la corrélation des attributs des entités permet de trouver des attributs synonymes d'un schéma relationnel et peut ainsi aider les utilisateurs débutants lors de la conception d'un schéma relationnel d'une base de données (Cafarella *et al.,* 2008). Les attributs peuvent également être utilisés pour enrichir automatiquement des bases de connaissances en les alimentant par de nouveaux triplets RDF (Gerber *et al.,* 2012).

3. Données exploitées dans le cadre de l'évaluation

Comme nous l'avons vu dans la section 1, dans le cadre de la recherche d'attributs d'une entité, et afin de savoir si le Web de Données permet d'ajouter de l'information pertinente par rapport aux informations du Web classique, nous avons mis en place une évaluation dans le but d'étudier les attributs provenant de ces deux sources. Pour

Figure 3. *Exemple de table relationnelle du Web*

le Web classique, nous avons utilisé la méthode présentée dans (Kopliku *et al.*, 2011) basée sur les tables relationnelles du Web, et pour le Web de Données nous avons exploité quelques datasets du projet *Linking Open Data*.

3.1. *Données issues du Web classique*

Étant donnée une entité, la méthode utilisée pour extraire les attributs pertinents à partir du Web se base sur les tables HTML relationnelles. Elle comporte 2 étapes (Kopliku *et al.*, 2011) :

– la recherche d'un ensemble de tables potentiellement pertinentes se trouvant dans les 50 premiers documents renvoyés par le moteur de recherche *Bing*[9] en réponse à la requête entité.

– l'application de filtres sur l'ensemble des tables trouvées afin d'éliminer les tables non relationnelles (comme les tables de mise en page) et les attributs non pertinents. Ces filtres sont implémentés en se basant sur une classification supervisée utilisant des paramètres comme la dimension de la table, des statistiques sur la longueur des lignes et des colonnes, la nature des valeurs des cellules (numérique, chaîne de caractère, ou date), des statistiques sur le nombre de caractères de ces valeurs, etc.

9. http ://www.bing.com/

3.2. *Données issues du Web de Données*

Le Projet *Linking Open Data* est une mise en pratique du concept de Web de Données. En septembre 2011, ce projet contenait 295 datasets et 31 milliards de triplets RDF. Les datasets sont répartis en 7 domaines : *Cross-Domain*, *Publications*, *Geographic*, *Media*, *Gouvernment*, *Life Sciences* et *User-Generated Content*. Etant donnée l'immense quantité d'information qu'ils contiennent et afin d'alléger la quantité d'attributs à traiter dans notre évaluation, nous nous sommes limités à certains datasets que nous avons sélectionnés selon certains critères :

– Nous avons éliminé les datasets qui n'offrent pas de terminal SPARQL en ligne ou une API pour l'accès aux informations.

– Nous avons également éliminé les datasets ne correspondant pas à notre tâche de recherche d'attributs, comme les datasets de vocabulaire (WorldNet, Yago, ...).

– Nous souhaitons conserver les datasets les plus populaires (car probablement les plus "importants"). Par analogie au principe du PageRank (Page *et al.*, 1998), un dataset est populaire s'il est bien référencé par les autres datasets.

– Pour alléger la quantité d'information à juger tout en essayant de diversifier les thèmes (publications, musiques, films, pays, ...), nous nous limitons pour chacun des 7 domaines définis dans le projet *Linking Open Data* à deux datasets (au maximum) n'ayant pas le même thème.

En appliquant ces heuristiques de sélection nous avons obtenu 11 datasets :

– 2 datasets génériques traitant divers thèmes (appartenant au domaine *cross-domain*) : *DBpedia* et *Freebase*.

– 9 datasets spécifiques : du domaine *Life Science* les datasets : *Drugbank* et *Diseasome*, du domaine *Government* les datasets : *Ordnance Survery* et *Statistics data.gov.uk*, du domaine *Geographic* les datasets *Geo Species* et *World Factbook*, du domaine *Publications* le dataset : *DBLP*, et du domaine *Media* les datasets : *BBC Music* et *Linked Movies DB*.

Étant donnée une entité e et 11 datasets de triplets RDF, nous voulons extraire tous les attributs de e. Pour cela nous avons utilisé des requêtes SPARQL qui ont la forme suivante :

```
1. SELECT ?predicat ?objet
2. FROM < dataset URI >
3. WHERE
4. {
5.     <?s> rdfs:label  e .
6.     <?s> ?predicat  ?objet.
7. }
```

Plus précisément, notre méthode comporte 3 étapes :

– La première étape consiste à chercher l'URI de l'entité *e* pour chaque dataset (ligne 5 de la requête ci-dessus). Le délai d'interrogation étant trop long pour *DBpedia*

et *Freebase* à cause de leur volume, nous avons pour ces datasets cherché l'URI de l'entité *e* d'une autre façon :

- Pour *DBpedia*, l'URI d'une entité *e* a la forme : *http ://dbpedia.com/resource/**e_name***. *e_name* désigne l'identifiant de *e*. Nous avons utilisé le moteur de recherche *Bing* et la requête (*site :en.wikipedia.org e"*) pour chercher l'identifiant *e_name* comme illustré sur la figure 4.

- De manière identique, pour *Freebase* nous avons également utilisé *Bing* et la requête (*site :freebase.com* e").

– La deuxième étape consiste à récupérer pour l'entité e, représentée par la variable *?s*, tous les prédicats et les objets (lignes 6 de la requête ci-dessus).

– La troisième étape consiste à extraire les noms d'attributs (à partir de la variable $?predicat$) et les valeurs correspondantes (à partir de la variable $?objet$) comme illustré dans la figure4.

predicat	objet
http://dbpedia.org/property/**capital**	http://dbpedia.org/resource/**Paris**
http://dbpedia.org/property/**currency**	"**Euro, CFP franc**"@en
http://dbpedia.org/property/**nationalAnthem**	""**La Marseillaise**""@en
http://dbpedia.org/property/**leaderName**	http://dbpedia.org/resource/**Fran%C3%A7ois_Hollande**
http://dbpedia.org/property/**callingCode**	"**33**"^^<http://www.w3.org/2001/XMLSchema#int>
http://dbpedia.org/ontology/**areaTotal**	**6.74842e+11**
http://dbpedia.org/ontology/**language**	http://dbpedia.org/resource/**French_language**
http://dbpedia.org/ontology/**capital**	http://dbpedia.org/resource/**Paris**
http://dbpedia.org/ontology/**anthem**	http://dbpedia.org/resource/**La_Marseillaise**
...	...

Figure 4. *Extrait des résultats de DBpedia pour la requête entité 'France'*

4. Mise en place de l'évaluation

Afin de connaître le potentiel de chacune des sources dans la tâche de recherche d'attributs d'une entité, nous avons mis en place une évaluation utilisateur qui consiste à évaluer les attributs d'un ensemble de requêtes (entités) choisies. Nous commençons par présenter ces requêtes, et nous détaillons ensuite le protocole de l'évaluation.

4.1. *Requêtes de l'évaluation*

Nous avons sélectionné 57 requêtes que nous espérons représentatives des requêtes entité. Pour ce faire, nous avons tout d'abord sélectionné 19 classes d'entités dont :

– 9 classes **spécifiques** qui coïncident avec les domaines des datasets spécifiques choisis à partir du projet Linking Open Data. Ces classes sont : *Diseases*, *Countries*, *Drugs*, *Films*, *Non-metropolitan counties*, *Songs*, *Articles*, *English Cities* et *Species*.

– 10 autres classes génériques dont

- 5 classes **génériques encyclopédiques** qui regrouperont des entités pouvant être décrites par une encyclopédie : *National Leaders*, *Places*, *Artists*, *Companies*, *Organisations*.

- 5 classes **génériques non encyclopédiques** qui regrouperont des entités qui ne sont pas pleinement décrites par une encyclopédie : *SLR Cameras*, *Software*, *Laptops*, *Mobiles phones*, *Programmable calculators*.

Pour chacune de ces classes, nous avons choisi aléatoirement 3 entités. Plusieurs membres de notre équipe ont proposé 5 entités pour chaque classe. Nous avons choisi automatiquement les requêtes proposées plus d'une fois, et pour le reste nous avons fait un tirage au sort pour atteindre une limite de 3 requêtes par classe.

Nous voulons savoir à travers cette classification des requêtes si les sources que nous utilisons pourraient répondre différemment aux différents types de requêtes. Les 57 requêtes sélectionnées sont disponibles à l'adresse : *http ://www.irit.fr/entityEval/query.jsp*.

4.2. *Procédure d'évaluation*

Nous avons mis en place une interface (figure 5) afin d'évaluer les 5783 attributs renvoyés au total par les deux sources pour les 57 requêtes. 14 volontaires ont participé à cette évaluation, chacun a évalué entre 3 et 5 requêtes. Le temps d'évaluation moyen mis par requête est de 20 minutes. Notre but étant d'évaluer le plus de requêtes possibles sur de nombreux datasets, chaque requête n'a été jugée que par un seul évaluateur.

Pour chacune des entités servant de requête, nous indiquons sur l'interface la classe correspondante. Ensuite, nous présentons tous les attributs un par un sans afficher

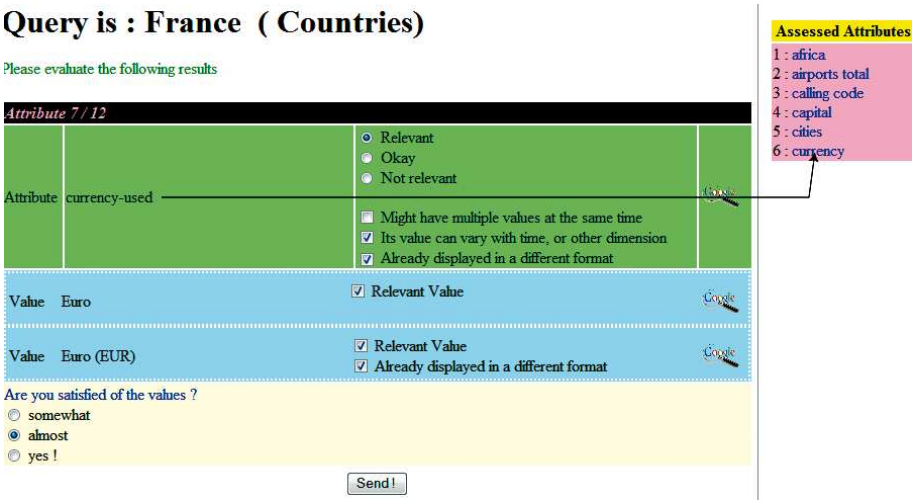

Figure 5. *Interface d'évaluation*

sa/ses source(s). Le travail d'un évaluateur consiste à prendre connaissance de la requête, évaluer le nom d'attribut, puis évaluer ses valeurs. Pour l'évaluation du nom d'attribut, l'utilisateur juge tout d'abord sa pertinence par rapport à la requête en cochant la case correspondante (*Relevant*, *Okay* ou *Not Relevant*). Dans le cas où il est pertinent ou moyennement pertinent, nous lui posons trois autres questions :

– L'attribut est-il multivalué, c'est à dire doit-il avoir plusieurs valeurs en même temps ? par exemple le nom d'attribut *'villes : Paris, Toulouse, Lyon, ...'* pour la requête *'France (Pays)'*.

– L'attribut peut-il dépendre d'autres dimensions (temps, localisation, température, ...) ? par exemple le nom d'attribut *'président :François-Hollande'* pour la requête *'France (Pays)'* dépend de la dimension temporelle.

– Le nom d'attribut est-il redondant, autrement dit, a-t-il déjà été affiché sous une autre forme ? Par exemple, l'attribut *"currency-used"* en cours d'évaluation sur la figure 5 a été déjà affiché à l'évaluateur sous une autre forme *"currency"*.

Si le nom d'attribut est pertinent, l'utilisateur procède à l'évaluation de ses valeurs. L'évaluation des valeurs consiste à cocher les valeurs pertinentes et indiquer si une valeur a été déjà affichée sous une autre forme pour le même nom d'attribut (exemple la valeur *'Euro (EUR)'* a été déjà affichée sous une autre forme : *'Euro'* pour le nom d'attribut *'currency-used'*).
Si un attribut est jugé multivalué et/ou dépendant d'autres dimensions, nous demandons à l'utilisateur d'exprimer sa satisfaction sur l'ensemble des valeurs proposées.
Enfin, à la fin de chaque requête, nous demandons à l'utilisateur d'exprimer sa sa-

	Web de Données	Tables HTML	Intersection	Toutes
Spéc (27)	$50,33$	$23,51$	$5,03$	$68,81$
Gén Ency (15)	$33,60$	$18,46$	$5,00$	$47,06$
Gén Non Ency (15)	$5,13$	$22,06$	$1,33$	$24,53$
Toutes (57)	$34,03$	$21,45$	$4,05$	$51,43$

Tableau 1. *Nombre moyen de noms d'attributs pertinents par source et par type de requête*

tisfaction sur l'ensemble des attributs proposés (non satisfait, peu satisfait, presque satisfait, très satisfait).

5. Analyses des résultats de l'évaluation

Dans cette section nous analysons les résultats obtenus. Nous commençons par analyser l'apport des deux sources en attributs pertinents. Ensuite, nous analysons la précision de ces sources. Puis, nous analysons les satisfactions des utilisateurs, et nous finissons par discuter le problème de la redondance de données.

5.1. *Apport des sources en attributs pertinents*

Pour analyser l'apport des sources en noms d'attributs pertinents, nous illustrons dans le tableau 1 le nombre moyen de noms d'attributs pertinents renvoyés par source et par type de requête (spécifique, générique encyclopédique, générique non encyclopédique).

La première constatation est que les tables relationnelles du Web permettent de répondre aux requêtes quelque soit leur type, avec un nombre moyen de noms d'attributs presque similaire (entre 18 et 23 noms d'attributs par requête). Ceci montre bien que les tables relationnelles du Web ont une bonne couverture des noms d'attributs d'entités quelque soit leur classe.

Pour le Web de Données, le nombre moyen de noms d'attributs renvoyés est de 34. Cependant, ce nombre varie selon le type de requête. En effet, il est élevé pour les requêtes spécifiques (50), à la moyenne pour les requêtes génériques encyclopédiques (33) et faible pour les requêtes génériques non encyclopédiques (5). Ceci montre que la couverture du Web de Données varie selon le domaine. Le fait qu'une requête coïncide avec un domaine spécifique du Web de Données permet d'avoir plus d'attributs pertinents. Nous constatons aussi que pour les requêtes génériques non ou peu décrites par une encyclopédie, la couverture du Web de Données est faible (en moyenne 5 attributs par requête), ce qui peut être expliqué par le fait que la population des datasets du Web de Données se base généralement sur des extractions automatiques à partir

	Web de Données	Tables HTML	Intersection	Deux sources
Nbre d'attributs	2859	3203	279	5783
Nbre d'attributs pertinents	1940	1223	231	2932
Précision	$67,85\%$	$38,18\%$	$82,79\%$	$50,70\%$

Tableau 2. *Précision des noms d'attributs*

d'encyclopédies, comme le cas *DBpedia* qui est construite à partir des info-box des articles de *Wikipédia*.

En comparant nos sources en terme de nombre moyen de noms d'attributs renvoyés par type de requête, nous voyons que le Web de Données répond mieux que les tables relationnelles du Web pour les requêtes spécifiques et génériques encyclopédiques. L'inverse est constaté pour les requêtes génériques non encyclopédiques.

Le Web de Données et les tables relationnelles peuvent renvoyer les mêmes noms d'attributs avec la même syntaxe. Le nombre d'intersections est faible, mais nous n'avons pas pris en compte les relations de synonymie possibles entre les termes.

5.2. *Précision*

Dans cette section nous analysons nos deux sources en terme de précision. Nous commençons par analyser la précision générale des noms d'attributs, puis nous détaillons la précision des datasets du Web de Données, et la précision des valeurs d'attributs.

5.2.1. *Précision générale des nom d'attributs*

Le tableau 2 illustre la précision des noms d'attributs issus des deux sources. Les noms d'attributs issus du Web de Données sont plus précis que ceux issus des tables HTML. Nous pouvons expliquer ceci par le fait que les méthodes d'extraction d'attributs à partir du Web Classique sont plus difficiles à mettre en place, ce qui génère beaucoup de bruit dans les résultats. Au contraire, l'extraction des attributs à partir du Web de Données est plus simple en se basant principalement sur le langage SPARQL.

5.2.2. *Précision des attributs du Web de Données*

Le tableau 3 montre que la précision des noms d'attributs du Web de Données diffère d'un dataset à un autre. Cette précision est faible si un dataset utilise beaucoup de prédicats utilitaires comme :

– les prédicats de catégorisation, ex. *subject*, *topic*, *type*, etc.

Dataset	**Précision**
Freebase	$76,04\%$
DBpedia	$65,98\%$
Diseasome	$57,14\%$
Factbook	$93,36\%$
Drugbank	$47,78\%$
BBC music	$43,86\%$
Ordnance survey	$59,09\%$
Linked MDB	$82,09\%$
Statistics data.gov.uk	$74,42\%$
DBLP	100%
Geo Species	$35,48\%$

Tableau 3. *Précision des attributs par source*

– les prédicats d'organisation (ou de mise en forme), ex. *align*, *width*, *after*, *before*, *activeYearStartDate*, *filename*, etc.

– les prédicats de similarité de sujet, ex. *sameas*, *seealso*, etc.

– les prédicats d'inclusion, ex. *contains*, *organizationWith this scope*, *people born here*, *has_part*, etc.

– les prédicats de description, ex. *subject*, *alias*, *synopsis*, *has_sentences*, etc.

Ces prédicats peuvent être utiles dans la procédure de recherche, mais ils ne doivent pas être transformés en noms d'attributs affichés dans les résultats.

Attributs issus d'une ontologie :

Les datasets du Web de Données peuvent s'appuyer sur une ou plusieurs ontologies. Certains noms d'attributs renvoyés sont issus de ces ontologies. Nous analysons ici la précision des attributs "ontologiques" pour le dataset *DBpedia*. Cette base de connaissance a renvoyé pour les 57 requêtes 1399 noms d'attributs (408 noms d'attributs ontologiques et 991 noms d'attributs non ontologiques) avec une précision générale de 66%. En analysant que les nom d'attributs ontologiques, nous avons trouvé que leur précision est de 84%. Ceci montre la bonne qualité des attributs issus d'ontologies. Cette constatation pourrait être exploitée dans le cas où nous voudrons trier les attributs renvoyés en donnant plus de poids pour les attributs issus d'une ontologie.

5.2.3. *Précision des valeurs d'attributs*

Pour les 1940 noms d'attributs pertinents issus du Web de Données, nous avons 1741 noms d'attributs avec au moins une bonne valeur, soit une précision de $89,74\%$. Les tables relationnelles du Web ont renvoyé 1223 noms d'attributs pertinents dont 898 avec au moins une bonne valeur, soit une précision de $73,42\%$.
Ces deux sources permettent de renvoyer des attributs pertinents avec des valeurs gé-

néralement précises. Mais la précision des valeurs issues du Web de Données reste meilleure que celle des valeurs issues des tables relationnelles.

5.3. *Satisfaction des utilisateurs*

Dans cette section, nous analysons la satisfaction générale des utilisateurs pour les attributs renvoyés. Ensuite, nous analysons leur satisfaction pour les attributs multivalués, puis pour les attributs dépendant d'autres dimensions.

5.3.1. *Satisfaction générale*

A la fin de chaque requête, nous avons demandé à l'évaluateur d'exprimer sa satisfaction de l'ensemble des attributs renvoyés. Généralement, les utilisateurs sont satisfaits ($47/57$).Cette satisfaction est plus importante pour les requêtes spécifiques ($25/27$) et moins importante pour les requêtes non encyclopédiques ($10/15$).

5.3.2. *Satisfaction pour les attributs multivalués*

Selon les résultats de l'évaluation, 33% des attributs renvoyés ont été jugés comme multivalués, c'est à dire qu'ils doivent avoir plus d'une seule valeur. Nous avons analysé la satisfaction des utilisateurs sur les attributs multivalués en fonction du nombre de valeurs proposées. Nous avons obtenu la courbe représentée dans la figure 6.

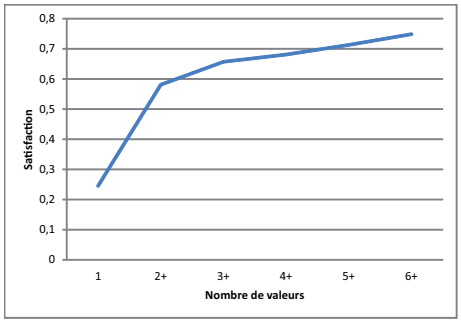

Figure 6. *Satisfaction pour les attributs multivalués en fonction du nombre de valeurs proposées*

Nous voyons clairement que la satisfaction des utilisateurs pour les valeurs d'attributs multivalués dépend du nombre de valeurs présentées : plus le nombre de valeurs présentées est grand, plus la satisfaction augmente.

5.3.3. *Satisfaction pour les attributs dépendants d'autres dimensions*

Les résultats de l'évaluation ont montré que 22% des attributs renvoyés dépendent d'une autre dimension. La courbe de la figue 7 présente la satisfaction des utilisateurs

en fonction du nombre de valeurs proposées (pour les attributs dépendant d'autres dimensions issus du Web de Données).

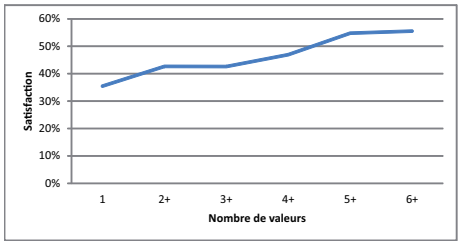

Figure 7. *Satisfaction pour les attributs dépendant d'une autre dimension en fonction du nombre de valeurs proposées*

La satisfaction s'améliore si on propose plus de valeurs, mais elle reste un peu faible (inférieure à 60%). Ceci est dû probablement à la manière de représenter ce type d'attribut. Par exemple un attribut dépendant de la dimension temporelle, l'idéal serait peut être de représenter l'évolution de ses valeurs selon un axe de temps.

5.4. *Redondance de données*

D'après les résultats de l'évaluation, 347 noms d'attributs sont retournés avec une même syntaxe par plus d'une source (ou dataset). 289 nom d'attributs parmi ces 347 noms d'attributs ont été jugés pertinents par les évaluateurs. Autrement dit, un nom d'attribut issu de plus d'une source a une probabilité de 83% d'être pertinent. La redondance d'un nom d'attribut peut donc refléter sa pertinence.

Nous pouvons aussi trouver des noms d'attributs redondants mais avec des syntaxes différentes, il s'agit de noms synonymes comme les noms d'attributs *currency* et *currency-used* dans la figure 5. D'après les résultats de l'évaluation, nous avons 575 synonymies, soit 19% des noms d'attributs pertinents retournés. Un système de recherche d'attributs doit détecter la redondance d'attributs pour deux raisons, d'une part, pour éviter la duplication des données présentées à l'utilisateur, et d'autre part, pour l'utiliser probablement comme critère pour renforcer les scores de tri des attributs redondants.

6. Conclusion et perspectives

Cet article s'inscrit dans le cadre de recherche d'attributs pour des requêtes de type entité. Nous nous sommes intéressés à récupérer des attributs d'une entité nommée, à partir de deux sources différentes : les tables relationnelles du Web et le Web de Données. Si la première source a été déjà utilisée dans plusieurs approches de l'état

de l'art, le Web de données reste sous-utilisé à notre connaissance.
Afin de savoir le potentiel de chacune de ces sources, nous avons mis en place une évaluation utilisateur. Nous avons choisi 57 requêtes appartenant à 19 classes différentes. Après avoir exécuté ces requêtes tout en exploitant le Web de données et les tables relationnelles du Web, nous avons demandé aux évaluateurs de juger les attributs obtenus. Nous avons séparé l'évaluation des noms d'attributs et celle des valeurs d'attributs.

Les résultats d'évaluation ont montré que les tables relationnelles du Web sont très importantes pour répondre à toutes les requêtes de type entité quelle que soit leur classe, mais leur précision reste faible par rapport au Web de Données. Ce dernier a une couverture un peu mois large puisqu'il ne répond pas suffisamment aux requêtes entité non encyclopédiques, mais sa précision est meilleure que celle des tables relationnelles. Nous déduisons donc l'utilité de combiner ces deux sources pour cette tâche de recherche d'attributs d'une requête entité.

Notre évaluation a également soulevé des problèmes qu'un système de recherche d'information devra résoudre lorsqu'il cherche et trie des attributs liés à une entité :

– Un bon système de recherche d'attributs doit détecter les redondances d'attributs, d'une part pour les éliminer lors de la présentation des résultats, et d'autre part pour s'en servir probablement comme critère pour renforcer les scores de tri des attributs redondants.

– Afin de satisfaire au mieux l'utilisateur d'un tel système de recherche d'information, les valeurs des attributs multivalués se doivent d'être le plus exhaustives possible. De même, pour les attributs dont la valeur dépend d'une autre dimension, la détection de cette dimension semble cruciale.

Malgré ces premiers résultats intéressants, notre évaluation pourrait être complétée :

– En effet, notre protocole d'évaluation ne nous permet pas d'évaluer l'exhaustivité par source des valeurs des attributs multivalués : l'utilisateur juge l'exhaustivité de toutes les valeurs issues de toutes les sources. Par exemple, nous considérons la requête *France* de la classe *Pays*, et l'attribut multivalué *Villes*. Supposons qu'une source $S1$ nous retourne toutes les villes existantes en France, alors que la source $S2$ nous retourne que quelques villes. La satisfaction de l'utilisateur sur les valeurs proposées varie dans ce cas selon la source, mais nous ne sommes pas capable de le détecter. Cette détection entraînerait une interface d'évaluation bien plus complexe que celle mise en place (voir même une deuxième interface), c'est pour cela que nous avons décidé à priori de ne pas faire cette analyse ;

– concernant l'extraction des noms d'attributs à partir des prédicats des triplets RDF, nous pourrions utiliser leurs labels et non pas une partie de l'URI du prédicat, ce qui rendrait les noms d'attributs plus précis.

Plusieurs perspectives s'ouvrent enfin pour ces travaux. Nous souhaitons construire un système de recherche d'information répondant aux requêtes de type entité, combinant le grand nombre de datasets du Web de Données, et le Web classique.

Ce système devra trier les attributs par ordre de pertinence, éviter la redondance, traiter les attributs multi-valués et dépendant d'autres dimensions. Dans notre cas, le tri des attributs en réponse à une requête reste encore un problème ouvert.
Nous aimerions aussi analyser si les réponses du Web classique sont plus "fraîches" que celles du Web de Données. Si cela s'avère vrai, nos techniques pourraient permettre de mettre à jour automatiquement certaines parties du Web de Données.

7. Bibliographie

Bellare K., Talukdar P., Kumaran G., Pereira F., Liberman M., McCallum A., Dredze M., « LightlySupervised Attribute Extraction for Web Search », *Proceedings of Machine Learning for Web Search Workshop, NIPS*, 2007.

Bizer C., Heath T., Berners-Lee T., « Linked Data The Story So Far », *International Journal on Semantic Web and Information Systems*, vol. 5, n° 3, p. 1-22, 2009.

Bizer C., Heath T., Idehen K., Berners-Lee T., « Linked data on the web (LDOW2008) », *Proceedings of the 17th international conference on World Wide Web*, WWW '08, ACM, New York, NY, USA, p. 1265-1266, 2008.

Cafarella M. J., Halevy A., Wang D. Z., Wu E., Zhang Y., « WebTables : exploring the power of tables on the web », *Proc. VLDB Endow.*, vol. 1, n° 1, p. 538-549, August, 2008.

Chen H.-H., Tsai S.-C., Tsai J.-H., « Mining tables from large scale HTML texts », *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, Stroudsburg, PA, USA, p. 166-172, 2000.

Chen P. P.-S., « The entity-relationship model-toward a unified view of data », *ACM Trans. Database Syst.*, vol. 1, n° 1, p. 9-36, March, 1976.

Gerber D., Ngomo A.-C. N., « Extracting Multilingual Natural-Language Patterns for RDF Predicates. », *EKAW*, vol. 7603 of *Lecture Notes in Computer Science*, p. 87-96, 2012.

Kopliku A., Pinel-Sauvagnat K., Boughanem M., « Attribute retrieval from relational web tables », *Proceedings of the 18th international conference on String processing and information retrieval*, SPIRE'11, Berlin, Heidelberg, p. 117-128, 2011.

Krichen I., Kopliku A., Pinel-Sauvagnat K., Boughanem M., « Une approche de recherche d'attributs pertinents pour l'agrégation d'information. », *INFORSID*, p. 385-400, 2011.

Page L., Brin S., Motwani R., Winograd T., « The PageRank citation ranking : Bringing order to the Web », *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, p. 161-172, 1998.

Pasca M., Van Durme B., « What you seek is what you get : extraction of class attributes from query logs », *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, San Francisco, CA, USA, p. 2832-2837, 2007.

Pound J., Mika P., Zaragoza H., « Ad-hoc object retrieval in the web of data », *Proceedings of the 19th international conference on World wide web*, WWW '10, New York, NY, USA, p. 771-780, 2010.

Tokunaga K., Kazama J., Torisawa K., « Automatic discovery of attribute words from web documents », *Proceedings of the Second international joint conference on Natural Language Processing*, IJCNLP'05, Berlin, Heidelberg, p. 106-118, 2005.