État des lieux de l'utilisation de OWL2 : Analyse et proposition pour capturer les utilisations de la sémantique OWL2 dans les graphes de connaissances RDF

Pierre-Henri Paris*,Fayçal Hamdi*
Samira Si-said Cherfi*

*Conservatoire National des Arts et Métiers
Laboratoire CEDRIC, équipe ISID
Paris, France
https://cedric.cnam.fr/index.php/labo/Isid
pierre-henri.paris@upmc.fr, faycal.hamdi@cnam.fr, samira.cherfi@cnam.fr

Résumé. Le nombre et la taille des graphes de connaissances RDF sont en constante augmentation. Par conséquent, le traitement des données pour des agents (automatisés ou humains) devient de plus en plus difficile. Si plusieurs outils peuvent être utilisés pour une tâche donnée, mais qu'ils dépendent chacun à des degrés divers de la sémantique disponible dans le graphe de connaissances, alors il est important d'avoir un aperçu en amont du graphe pour sélectionner le meilleur outil pour cette tâche. Nous avons mené, à grande échelle, une étude approfondie pour vérifier la présence de sémantique dans les graphes de connaissances publiés actuellement dans le Web de données (*Linked Data*). Bien que certains graphes de connaissances utilisent la sémantique OWL 2, beaucoup ne le font pas ou partiellement. Nous proposons donc une approche qui, en se basant sur des statistiques, instancie une ontologie facilitant la sélection de l'outil le mieux adapté à une tâche donnée en fonction de l'utilisation de la sémantique OWL 2.

1 Introduction

Avec l'augmentation du nombre et de la taille des graphes de connaissances RDF, la difficulté pour requêter ou utiliser ces données s'accroît. Pour une tâche donnée, plusieurs types d'approches peuvent être envisagés. Certaines approches reposant principalement sur la sémantique disponible dans les graphes, d'autres au contraire ne l'utilisant que peu ou pas du tout. Bien entendu, entre ces deux extrêmes, des approches peuvent tirer parti de la sémantique, sans toutefois se reposer totalement sur cette dernière. Par exemple, si la tâche est d'interconnecter plusieurs graphes de connaissances entre eux, les approches peuvent utiliser une combinaison de techniques telles que les statistiques, d'autres graphes de connaissances externes, la sémantique ou encore des algorithmes de partitionnement de données. Par ailleurs, des approches se reposant principalement sur la sémantique peuvent surpasser d'autres types d'approches si la sémantique est très présente dans le graphe de connaissances. Mais si la

sémantique est absente, les résultats risquent fort de ne pas être conformes aux attentes de l'utilisateur. Il est donc souvent nécessaire de conduire une première étude exploratoire du graphe de connaissances afin de savoir quel outil conviendra le mieux pour la tâche donnée. Cette étude permet ainsi de comprendre ce que les données peuvent avoir à offrir. Malheureusement, cette étape exploratoire est très consommatrice en termes de temps, d'autant plus si la documentation accompagnant le graphe de connaissances est absente ou peu éclairante. Plusieurs vocabulaires ou ontologies ont été proposés afin de fournir à l'utilisateur un aperçu des données contenues dans le graphe de connaissances. Par exemple, Dublin Core [1] de Weibel et al. (1998), Creative Commons Rights Expression Language [2], Data Catalog Vocabulary [3], ou VoID [4] de Alexander et al. (2009) permettent de décrire les graphes de connaissances, mais ne donne pas la possibilité d'exprimer quels sont les éléments de OWL 2 utilisés.

Dans ce papier, nous avons, dans un premier temps, mené une étude sur l'utilisation de OWL 2 pour décrire les graphes de connaissances des données liées ouvertes. Cette étude est à grande échelle puisqu'elle porte sur plus de 600 mille graphes de connaissances. Elle a pour objectif de confirmer ou d'infirmer plusieurs résultats sur la manière dont ces ontologies exprimées en OWL 2 sont utilisées dans les données liées ouvertes. Cette étude permet aussi d'apprécier le degré de variété dans l'utilisation des différentes fonctionnalités de OWL 2. Dans un second temps, nous avons conçu une ontologie permettant d'exprimer, pour un graphe de connaissances donné, quelles sont les fonctionnalités OWL 2 utilisées et dans quelles proportions. Ainsi, cette ontologie permet d'apporter directement au consommateur de données les informations nécessaires pour sélectionner en connaissance de cause l'outil approprié à la réalisation de sa tâche. Enfin, nous avons fourni les requêtes SPARQL permettant d'instancier l'ontologie pour un graphe de connaissances donné. Ainsi, les consommateurs de données pourront précisément connaître la manière et l'étendue de l'utilisation de OWL 2 dans le graphe de connaissances. Ceci, grâce à des statistiques agrégées à propos de tous les vocabulaires ou ontologies décris avec OWL 2 et qui sont utilisés dans le graphe de connaissance.

Dans la section suivante, nous décrivons les différents travaux ayant mesuré le degré d'utilisation de la sémantique OWL 2 dans les graphes de connaissances. Puis, nous présentons notre étude sur l'état actuel des données liées ouvertes. Ensuite, nous présentons l'ontologie qui permet de décrire l'utilisation de OWL 2 dans les graphes de connaissances et les requêtes SPARQL qui permettent de l'instancier pour un graphe de connaissances donné. Enfin, nous concluons et donnons quelques perspectives.

2 État de l'art

Dans cette section, nous présentons quelques travaux qui se concentrent, d'une manière ou d'une autre, sur l'étude de l'utilisation de la sémantique dans les graphes de connaissances des données liées ouvertes.

Dans d'Aquin et al. (2007), les auteurs ont analysé 25500 graphes de connaissances en termes d'expressivité. Bien qu'intéressante, cette étude est ancienne et porte sur un très petit nombre de graphes de connaissances.

1. http://www.dublincore.org/specifications/dublin-core/
2. https://creativecommons.org/ns
3. https://www.w3.org/TR/vocab-dcat/
4. https://www.w3.org/TR/void/

Jain et al. (2010) dénonce le manque d'expressivité des graphes de connaissances, c'est-à-dire que beaucoup de graphes de connaissances n'utilisent pas toutes les différentes fonctionnalités de OWL 2, loin s'en faut. De ce fait, nombre d'approches reposant sur les fonctionnalités les plus avancées de OWL 2 sont inutilisables en l'état sur ces graphes utilisant des ontologies peu expressives. Par exemple, sans propriétés qui pourraient être déclarées comme étant transitives mais qui ne sont pas décrites ainsi, il est plus compliqué de naviguer entre les données liées alors que ceci est censé être l'une des forces de ce type de graphe de connaissances. De plus, ce papier n'analyse que 70 graphes du *LOD Cloud*[5], ce qui est peu, eut égare à sa taille actuelle (1239 graphes).

Dans Hitzler et van Harmelen (2010), les auteurs mettent l'accent sur le fait que quelques éditeurs de données se concentrent uniquement sur la publication des données (c'est-à-dire des triplets) sans les annoter avec des ontologies partagées. Ceci réduit donc la possibilité de résonner sur ces données. Leur conclusion est que de manière générale, mis à part la propriété *owl:sameAs*, les fonctionnalités de OWL 2 sont peu utilisées. Toutefois, cette étude est plus un constat empirique qu'une étude systématique.

Hogan et al. (2010) et Polleres et al. (2010) affirment que la qualité des données liées ouvertes peut poser problème en raison d'un manque de définitions des propriétés et des classes, c'est-à-dire quand aucune description en OWL 2 ou RDFS n'est disponible. Alors que, par exemple, la définition de classes disjointes peut aider à détecter les inconsistances dans un graphe de connaissances. Hogan et al. (2010) porte sur 12,5 millions de triplets et a pour objectif de soulever les différents problèmes auxquels doit faire face le Web Sémantique. Par exemple, les auteurs relèvent le nombre de problèmes de déréférencement, de syntaxe au niveau RDF ou encore les problèmes d'inconsistances. Mais en raison de la petitesse de l'échantillon et de l'ancienneté de l'étude d'une part et d'autre part du manque de métriques pertinentes sur l'utilisation de la sémantique, cette étude n'apporte pas de réponse à nos questions.

Färber et al. (2016) propose d'investiguer la qualité de certains des plus connus des graphes de connaissances. Les auteurs proposent des statistiques basiques sur DBpedia, Freebase, OpenCyc, Wikidata et YAGO. Bien que n'étant pas une étude à large échelle de l'utilisation de la sémantique, certaines statistiques s'avèrent intéressantes (nombre de triplets, nombre de classes, nombre de relations, etc.), mais ne s'intéressent pas suffisamment à la sémantique exprimée par les ontologies fondées sur OWL 2.

Aucun des travaux cités ne propose d'étude complète sur l'utilisation de la sémantique OWL 2 dans les graphes de connaissances RDF avec des chiffres précis et à une telle échelle. Dans ce papier, nous proposons donc de recueillir de l'information sur l'utilisation à très grande échelle des fonctionnalités OWL 2. Nous fournissons également une ontologie décrivant ces fonctionnalités, et les requêtes SPARQL permettant de l'instancier.

3 Etat actuel des données ouvertes liées

Dans cette section, nous présentons les sources que nous avons utilisées afin de produire des résultats sur une plus grande échelle. Ensuite nous décrirons la méthodologie que nous avons employée pour mener à bien cette étude et nous présentons les différents résultats obtenus. **L'objectif de cette étude est d'avoir un aperçu, sur une large échelle et sur des données**

5. `https://lod-cloud.net/`

les plus récentes possible, de l'utilisation de la sémantique exprimée en OWL 2 dans les graphes de connaissances RDF.

Avant de poursuivre, nous proposons la définition suivante qui nous sera utile à la fois pour l'ontologie, mais aussi pour l'étude que nous allons entreprendre.

Definition 3.1. (**Fonctionnalité sémantique**) Une fonctionnalité sémantique est n'importe quel élément de OWL 2, c'est-à-dire toutes ses propriétés et classes comme $owl{:}Restriction$ or $owl{:}SymmetricProperty$.

Par exemple, $\langle{:}age\ a\ owl{:}FunctionalProperty\rangle$ et tous les triplets utilisant *:age* en tant que prédicat. Nous prendrons en considération à la fois les triplets définissant $:age$ et l'utilisation de cette propriété en tant que prédicat dans des triplets.
Ou encore, $\langle{:}someClasses\ a\ owl{:}AllDisjointClasses\rangle$ et tous les triplets de la forme $\langle ?x\ rdf{:}type\ {:}someClasses\rangle$ sont aussi considérés comme des triplets utilisant une fonctionnalité sémantique de OWL 2.

3.1 Sources

Dans cette section, nous allons présenter les différentes sources de données que nous avons envisagé d'utiliser pour notre étude sur l'utilisation de la sémantique exprimée sous forme d'ontologies OWL 2. Nous expliquerons pourquoi nous avons retenu certaines d'entre elles et écarté les autres. Notre objectif est de réunir des informations sur l'utilisation de la sémantique OWL 2 dans les graphes de connaissances RDF.

Le *LOD Cloud* permet d'avoir un aperçu visuel de l'étendue et du développement des graphes de connaissances RDF ces dernières années. *LOD Cloud* référence plus de 1000 graphes de connaissances.

LOD Laundromat [6] (Beek et al. (2014)) donne accès à plus de 650 mille graphes de connaissances au format HDT (Fernández et al. (2013) and Martínez-Prieto et al. (2012)). HDT est un format compressé contenant des triplets RDF et permettant d'effectuer des opérations de recherches sur ces triplets. Les données de ces graphes ont été nettoyées. Les erreurs de syntaxes, les duplications de données et les noeuds anonymes par exemple ont été supprimés ou traités. Chaque graphe est contenu dans un fichier unique ayant un identifiant unique, et des métadonnées sont aussi disponibles (provenance, nombre de triplets, date de traitement, etc.). Certains de ces graphes concernent différentes versions du même jeu de données, par exemple DBpedia-fr, DBpedia-en ou DBpedia 3.8. Les graphes de *LOD Cloud* sont contenus dans *LOD Laundromat*.

LOD-a-lot [7] est un service fournissant l'accès aux mêmes données que *LOD Laundromat*, mais sous la forme d'un graphe unique dans lequel l'ensemble des triplets des graphes de *LOD Laundromat* forment un seul et unique graphe de connaissances. Bien que très intéressant, le fait de fusionner tous les graphes en un seul ne nous permet plus d'effectuer une distinction entre les différents graphes originaux et donc l'analyse résultante aurait été amoindrie. C'est pourquoi nous n'avons pas retenu *LOD-a-lot* comme source de données pour notre étude.

Nous avons donc choisi d'utiliser *LOD Laundromat* pour son grand nombre de graphes de connaissances et leur sérialisation au format HDT.

6. `http://lodlaundromat.org/`
7. `http://lod-a-lot.lod.labs.vu.nl/`

Nom	Remarque
# de triplets	Taille du graphe de connaissances
# de sujets distincts	Couverture du graphe
# de sujets sans type explicite	Le type d'un sujet est une information très basique
# de prédicats distincts	Nécessaire pour mettre en perspective les chiffres suivants
# de prédicats sans domaine spécifié	Complétude de la définition des propriétés
# de prédicats sans codomaine spécifié	Complétude de la définition des propriétés
pour chaque fonctionnalité OWL 2	
# de triplets l'utilisant	
# de sujets distincts l'utilisant	

TAB. 1: Informations collectées pour chaque graphe de connaissances.

3.2 Collecte d'informations

Dans cette section, nous décrivons comment, à partir des sources vues dans la section précédente, nous avons collecté les informations. Pour un graphe donné, nous avons analysé les différentes classes et propriétés d'une part, c'est-à-dire leurs définitions en OWL 2, et d'autre part nous avons analysé leurs utilisations au sein de ce graphe de connaissances. Pour la première analyse, si jamais la définition d'une classe (respectivement d'une propriété) n'est pas présente explicitement dans le graphe, alors nous avons cherché à atteindre cette définition par déréférencement, c'est-à-dire en allant voir si l'URL renvoie bien à un document RDF contenant la définition recherchée. Ainsi, cela nous permet par exemple de savoir si telle propriété est définie en tant que propriété fonctionnelle et de connaître le nombre de fois où elle est utilisée. En effet, il arrive qu'une classe ou une propriété soit définie, mais ne soit pas du tout utilisée dans les données. De plus, pour chaque propriété, nous avons cherché tous ses éventuels parents : par exemple, si p_1 est définie en tant que sous propriété de p_2 et que cette dernière est définie en tant que propriété fonctionnelle, alors p_1 sera aussi fonctionnelle. Cela nous assure ainsi de ne pas oublier des propriétés qui auraient telle ou telle fonctionnalité OWL 2. Pour chaque graphe, nous avons calculé les chiffres présentés dans la Table 1.

Le logiciel développé pour cette occasion l'a été en Java et est disponible pour réplication sur GitHub [8],ainsi que tous les résultats de cette étude et quelques résultats complémentaires qui n'ont pu être présentés par soucis de concision.

3.3 Résultats Généraux

Grâce à *LOD Laundromat*, 647 858 graphes de connaissances ont été analysés (un fichier HDT représente un graphe). Ainsi, dans cet article, un graphe de connaissance RDF est une sérialisation d'un graphe exprimé à l'aide du modèle de graphe RDF, c'est-à-dire composé de triplets sujet-prédicat-valeur. Il contient des données (A-Box) et une ontologie (T-Box). Nous allons, dans un premier temps, présenter les résultats généraux. Ensuite, nous regarderons en détail les résultats pour les différentes fonctionnalités de OWL 2.

La première vue sur ces résultats est présentée dans la Table 2. La première colonne correspond au sélecteur, c'est-à-dire le filtre que nous avons appliqué pour sélectionner un sous ensemble des presque 650 mille graphes. Ce sélecteur permet de choisir tous les graphes, ou

8. `https://github.com/PHParis/sem_web_stats`

Sélecteur	# de graphes	Q1	Q2	Q3	% de graphes avec au moins une fonctionnalité OWL 2
TOUS	647,858	547	3146.5	38,580	1.53
avec sémantique	10,600	184	3952	93,891	100
TOP 100 (# triplets)	100	5,611,982	8,951,766	12,635,325	34

TAB. 2: Statistiques basiques par sélecteur en termes de nombre de sujets (1er quartile, médiane et 3e quartile) et pourcentage de graphes de connaissances utilisant OWL 2 (au moins une fonctionnalité OWL 2 doit être utilisée au moins une fois). Le sélecteur "TOUS" correspond à tous les graphes, le sélecteur "avec sémantique" correspond à tous les graphes utilisant au moins une fonctionnalité OWL 2, et enfin, le sélecteur "TOP 100" correspond aux 100 plus gros graphes en terme de nombre de triplets.

les graphes ayant au moins une fonctionnalité OWL 2, ou les 100 plus grands en termes de nombre de triplets. La seconde colonne indique le nombre de graphes sélectionnés par le sélecteur (le maximum étant bien entendu 647 858 graphes). Dans les trois colonnes suivantes, nous indiquons les 1er, 2e et 3e quartiles en fonction du nombre de sujets distincts afin de visualiser la répartition des graphes. Pour finir, la dernière colonne montre le pourcentage de graphes contenant au moins une classe ou propriété utilisant l'une des fonctionnalités de OWL 2. Nous observons dans la Table 2 que les grands graphes (ceux du top 100) utilisent beaucoup plus de sémantique que la moyenne. Les graphes utilisant de la sémantique OWL 2 varient beaucoup en termes de taille puisque les 25% les plus petits utilisent beaucoup moins de sujets que dans le cas général.

En outre, la Figure 1 montre le nombre de graphes par fonctionnalités OWL 2. L'axe horizontal est le nombre de fonctionnalités OWL 2 distinctes contenues dans un graphe, par exemple *owl:sameAs*, *owl:FunctionalProperty*. L'axe vertical montre le nombre de graphes sur une échelle logarithmique. Nous pouvons voir que la grande majorité des graphes de connaissances des données liées ouvertes contient très peu de fonctionnalités OWL 2. Seulement 719 graphes de connaissances utilisent plus de 5 fonctionnalités distinctes de OWL 2 et 9881 n'utilisent qu'une.

3.4 Résultats par fonctionnalité OWL 2

Dans cette section, nous nous concentrons sur chaque fonctionnalité OWL 2 selon trois perspectives différentes. Nous allons d'abord observer les types de propriétés (fonctionnelle, etc.). Nous verrons ensuite les différents types de classes offerts par OWL 2 et enfin nous observerons les propriétés de OWL 2 (*owl:sameAs*, *owl:inverseOf*, etc.). Pour ces perspectives, nous aurons besoin d'une définition sur l'usage des classes et propriétés OWL 2.

Definition 3.2. (**Usage des propriétés et classes de OWL 2**) Soit C une classe OWL 2 (resp. p une propriété OWL 2). Soit $\Omega = \{KG_i | i \in [1, N]\}$ un ensemble de graphes de connaissances.

L'**usage de la classe** (resp. l'**usage de la propriété**) C (resp. p), noté $CU_\Omega(C)$ (resp. $PU_\Omega(p)$), est la moyenne pondérée du nombre de sujets ayant C comme type RDF (resp. utilisant p en tant que prédicat). Les poids sont l'inverse du nombre total de sujets dans le graphe

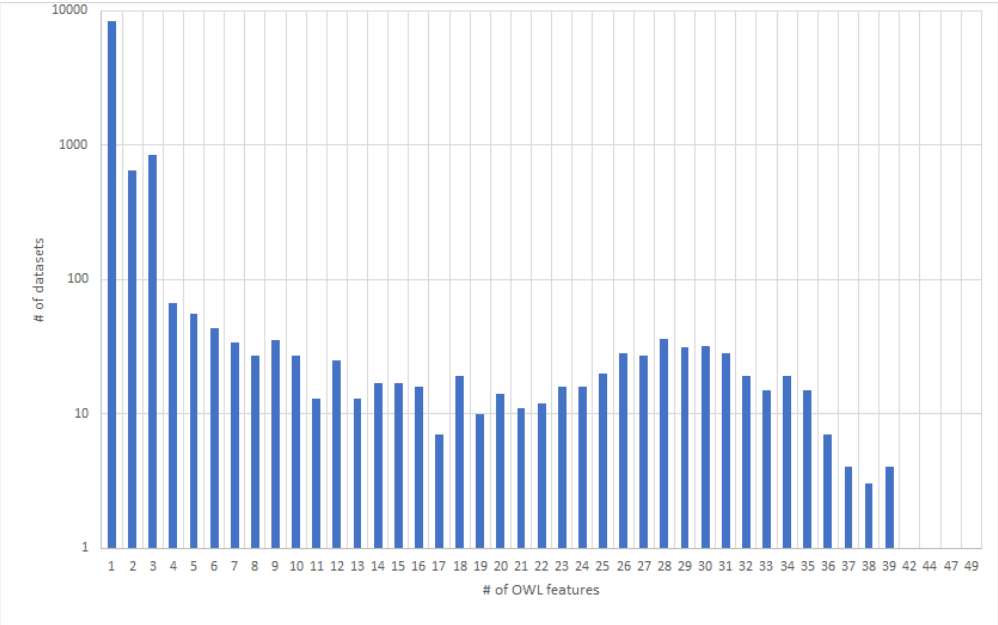

FIG. 1: Nombre de graphes de connaissances par fonctionnalités OWL 2.

de connaissances.

$$CU_{\Omega}(C) = \frac{\sum_{i=1}^{N} |Sub(KG_i, C)| \times \frac{1}{|Sub(KG_i)|}}{\sum_{i=1}^{N} \frac{1}{|Sub(KG_i)|}} \tag{1}$$

$$PU_{\Omega}(p) = \frac{\sum_{i=1}^{N} |Sub(KG_i, p)| \times \frac{1}{|Sub(KG_i)|}}{\sum_{i=1}^{N} \frac{1}{|Sub(KG_i)|}} \tag{2}$$

où $Sub(KG_i, C)$ donne tous les sujets distincts de type C dans KG_i, $Sub(KG_i, p)$ donne tous les sujets utilisant p dans KG_i et $Sub(KG_i)$ donne tous les sujets distincts dans KG_i.

La Définition 3.2 empêche les grands graphes de peser trop lourd sur la moyenne.

Par exemple, supposons que $C = owl{:}AllDisjointClasses$ est une classe OWL 2. Supposons aussi que $\Omega = \{KG_1, KG_2\}$, où KG_1 a 158 sujets distincts et 2 d'entre eux sont de type $owl{:}AllDisjointClasses$, et KG_2 a 357 sujets distincts et 9 d'entre eux sont de type $owl{:}AllDisjointClasses$. Alors $CU_{\{KG_1,KG_2\}}(owl{:}AllDisjointClasses) = \frac{2\times\frac{1}{158}+9\times\frac{1}{357}}{\frac{1}{158}+\frac{1}{357}} = 2,77$.

La Table 3 concerne les types des propriétés (par exemple une propriété qui serait définie comme étant fonctionnelle). Dans notre étude, un prédicat qui est une sous propriété d'une propriété fonctionnelle est aussi fonctionnelle. La seconde colonne montre le nombre de graphes utilisant cette propriété. La troisième colonne montre la moyenne pondérée du nombre de triplets utilisant la propriété du type considéré dans un graphe. Les deux dernières colonnes sont

Type	# de graphes	moyenne pondérée des triplets	moyenne pondérée des sujets	moyenne pondérée des prédicats
ObjectProperty	747	27.5	16.57	7.53
DatatypeProperty	685	31	19.43	5.79
FunctionalProperty	434	9	5.76	3.06
InverseFunctionalProperty	310	22.7	20.6	2.54
TransitiveProperty	396	2.84	2.63	2.4
SymmetricProperty	320	7	4.77	2.87
AsymmetricProperty	15	4.7	4.66	4.66
IrreflexiveProperty	21	1.66	1.65	1.65
ReflexiveProperty	16	1.32	1.32	1.32

TAB. 3: Analyse par type de propriété.

Classe	# de graphes	Utilisation de la classe
Class	1905	1.36
Restriction	520	10.3
DataRange	225	1.71
AllDifferent	213	2.35
NamedIndividual	62	10.8
AllDisjointClasses	50	2.09
NegativePropertyAssertion	27	279.76
Axiom	13	14.44
AllDisjointProperties	5	4.96

TAB. 4: Analyse par type de classe (voir Déf. 3.2).

similaires, mais pour les sujets et les prédicats. Par exemple, les propriétés inverses fonctionnelles sont trouvées dans 310 graphes. Parmi ces 310 graphes, on peut s'attendre à trouver en moyenne 2,54 définitions de telles propriétés qui sont utilisées sur 22,7 triplets avec 20,6 sujets différents. Comme nous pouvons le voir, certains prédicats sont très peu utilisés, comme les *owl:ReflexiveProperty* qui ne le sont que dans 16 graphes. Dans ces 16 graphes, très peu sont définis (1,28) et utilisés.

La Table 4 concerne la définition des classes en utilisant OWL 2. La deuxième colonne montre le nombre de graphes utilisant la classe donnée. La troisième colonne montre l'usage de la classe (voir Déf. 3.2). Par exemple, la classe *owl:Restriction* est utilisée dans seulement 520 graphes. Ces 520 graphes l'utilisant en moyenne 10,3 fois. Comme attendu, le type *owl:Class* est le plus utilisé au contraire de *owl:AllDisjointProperties* qui n'est pratiquement pas utilisé (5 graphes seulement).

La Table 5 montre l'usage des propriétés de OWL 2 dans les graphes des données liées ouvertes. La seconde colonne montre le nombre de graphes dans lesquels la propriété existe. La dernière colonne montre l'usage de cette propriété (voir Déf. 3.2). Par exemple, si un graphe utilise des fonctionnalités OWL 2, l'utilisateur peut s'attendre à trouver 5,41 *inverseOf* dans ce graphe.

Comme nous pouvons le voir, la propriété *owl:sameAs* est de loin la plus utilisée des pro-

Propriété	# de graphes	Utilisation des propriétés
sameAs	7708	10.30
unionOf	1256	1.02
inverseOf	548	5.41
onProperty	522	10.19
equivalentClass	492	5.11
disjointWith	470	2.49
allValuesFrom	425	7.92
someValuesFrom	413	10.07
cardinality	412	3.1
minCardinality	398	4.26
intersectionOf	394	5.26
maxCardinality	388	4.88
oneOf	364	1.99
hasValue	348	5.58
equivalentProperty	324	4.61
complementOf	315	1.82
distinctMembers	207	2.26
differentFrom	144	30.1
onClass	91	2.99

Propriété	# de graphes	Utilisation des propriétés
members	55	2.13
propertyChainAxiom	38	1.76
onDataRange	32	7.2
qualifiedCardinality	32	2.63
assertionProperty	27	279.76
sourceIndividual	27	279.76
targetIndividual	27	279.76
minQualifiedCardinality	25	4.24
disjointUnionOf	24	2.13
withRestrictions	13	1.04
annotatedTarget	12	14.83
maxQualifiedCardinality	12	4.06
annotatedProperty	10	12.71
annotatedSource	10	12.71
onDatatype	10	1.16
hasKey	8	1
propertyDisjointWith	5	1.08
hasSelf	2	1.88
datatypeComplementOf	1	1

TAB. 5: Analyse des propriétés OWL 2 (voir Déf. 3.2).

priétés de OWL 2, puisqu'elle est trouvée dans 6 fois plus de graphes que la seconde propriété la plus utilisée (*owl:unionOf*). De plus, *owl:sameAs*, lorsqu'elle est utilisée, elle l'est de manière intensive. En moyenne 10,3 triplets l'utilisent. Ceci correspond aux découvertes des travaux plus anciens, tel que Hitzler et van Harmelen (2010), montrant par la même occasion l'importance de *owl:sameAs* et l'inertie face au changement que l'on peut rencontrer sur ce type de graphe.

En conclusion, un très grand nombre de graphes de connaissances n'utilisent pas ou peu de sémantique sous forme de OWL 2 et quelques un en utilise beaucoup. De plus, de nombreuses fonctionnalités OWL 2 ne sont que très peu utilisées. Toutefois, plus un graphe est grand et plus il y a de chance qu'il utilise OWL 2.

4 Conception de l'ontologie

L'ontologie que nous proposons (disponible sur [9]) a pour but d'expliciter l'utilisation des classes et propriétés définies avec des éléments de OWL 2 et RDFS dans un graphe de connaissances. Basiquement, l'objectif est que l'utilisateur soit capable d'avoir facilement accès, par exemple, au nombre de propriétés qui sont transitives et à leur nombre d'utilisations dans le graphe.

VoID [10] de Alexander et al. (2009) est un vocabulaire qui permet de décrire un graphe de connaissances facilitant ainsi sa découverte et son utilisation. Il propose des statistiques très simples telles que le nombre de classes ou de triplets. Notre ontologie étend ce vocabulaire par une gestion des statistiques plus fines concernant l'utilisation des éléments de OWL 2 et

9. `http://cedric.cnam.fr/isid/ontologies/OntoSemStats.owl`
10. `https://www.w3.org/TR/void/`

RDFS. Ainsi, nous représentons un graphe de connaissances comme une instance de la classe $void{:}Dataset$. Cette instance pourra avoir autant de *:Stat* [11] qu'il utilisera de propriétés ou classes OWL 2 et RDFS. Chaque instance de *:Stat* a une et une seule instance *:SemanticFeature*. La propriété *:hasSemanticFeature* (voir Listing 1) permet de lier une instance de *:Stat* à sa *:SemanticFeature*. Les différents types du codomaine de *:hasSemanticFeature* sont disjoints deux à deux, permettant ainsi de détecter toute erreur dans l'instanciation de cette ontologie.

```
:hasSemanticFeature rdf:type owl:ObjectProperty , owl:FunctionalProperty ,
    owl:AsymmetricProperty , owl:IrreflexiveProperty ;  rdfs:domain :Stat ;
rdfs:comment "Specify which OWL 2 or RDFS semantic feature is the target
  of the given stat."@en ; rdfs:label "has semantic feature"@en .
```

Listing 1: Définition de la propriété hasSemanticFeature.

Pour ne pas changer de niveau de langage OWL 2 [12], nous ne pouvons utiliser de fonctionnalité OWL 2 en tant qu'objet d'un triplet. Par exemple, l'utilisation de *owl:FunctionalProperty* en tant qu'objet d'un triplet tel que $\langle{:}stat\ {:}hasSemanticFeature\ owl{:}FunctionalProperty\rangle$ entraînerait des problèmes d'indécidabilités [13] [14]. De ce fait, nous avons créé une sous classe de *:SemanticFeature* appelée *:FunctionalProperty* pour représenter les statistiques des propriétés fonctionnelles. Nous avons procédé ainsi pour chaque fonctionnalité OWL 2 et RDFS. Les différents axiomes de OWL 2 et RDFS peuvent impacter des propriétés, des classes ou des instances. Ainsi, nous avons choisi de faire en sorte, dans le design de notre ontologie, qu'elle reflète ceci. En fonction de son utilité, un axiome sera "rangé" dans une classe particulière. Par exemple, le Listing 2 montre la définition de *:PropertyType* (sous classe de *:SemanticFeature*) servant à représenter les différents types que peut avoir une propriété (symétrique, réflexive, etc.). Un autre exemple est la classe "PropertyRelation" qui regroupe entre autres les statistiques concernant *owl:propertyChainAxiom* ou *owl:inverseOf* qui sont des axiomes permettant de décrire la nature de la relation entre des propriétés. Toutes les définitions sont disponibles sur la page de l'ontologie.

```
:PropertyType rdf:type owl:Class ; rdfs:subClassOf :PropertyAxiom ;
  owl:disjointUnionOf ( :OwlAsymmetricProperty
    :OwlFunctionalProperty :OwlInverseFunctionalProperty
    :OwlIrreflexiveProperty :OwlReflexiveProperty
    :OwlSymmetricProperty :OwlTransitiveProperty) .
```

Listing 2: Définition de la classe Properties qui représente les différents types servant à définir une propriété.

Afin de fournir les statistiques concernant chaque fonctionnalité de OWL 2, nous avons créé deux propriétés : *:definitionCount* et *:usageCount*. La première sert à dire combien il y a de définition d'un axiome (par axemple le nombre de propriétés fonctionnelles) et la seconde sert à dire combien de fois l'axiom est utilisé (par exemple combien de triplets utilise une propriété fonctionnelle). Le Listing 3 montre la définition de la propriété *:usageCount* qui permet de dire, par exemple, que 3000 triplets utilisent une propriété fonctionnelle.

```
:usageCount rdf:type owl:DatatypeProperty ,
  owl:FunctionalProperty ;
```

11. *Les classes et les propriétés représentées sans préfixe appartiennent à notre ontologie*
12. https://www.w3.org/TR/owl2-profiles/#Computational_Properties
13. https://www.w3.org/2007/OWL/wiki/Profile_Explanations
14. https://www.w3.org/TR/owl2-profiles/

```
    rdfs:domain :Stat ;
    rdfs:range xsd:integer ;
    rdfs:comment "Number of usage of a semantic feature."@en ;
    rdfs:label "usage count"@en .
```

Listing 3: Définition de la propriété permettant de spécifier combien de propriétés d'un type donné sont utilisées concrètement.

Enfin, nous avons proposé les requêtes SPARQL nécessaires pour instancier l'ontologie. Le Listing 4 montre un exemple lorsque l'on souhaite instancier l'ontologie sur la propriété *owl:FunctionalProperty*.

```
CONSTRUCT {
    _:b0 a void:Dataset ; :hasStat [
        a :Stat ; :hasSemanticFeature :OwlFunctionalProperty ;
        :definitionCount ?definitionsCount ; :usageCount ?triples ]}
WHERE {
    {SELECT (COUNT (DISTINCT ?p) as ?definitionsCount)
        WHERE {
            { ?p a owl:FunctionalProperty . }
            UNION
            { ?p2 a owl:FunctionalProperty .
              ?p (rdfs:subPropertyOf|owl:equivalentProperty)+ ?p2 . }}}
    {SELECT (COUNT (*) as ?triples)
        WHERE {
            { SELECT DISTINCT ?p WHERE {
              { ?p a owl:FunctionalProperty . }
              UNION
              { ?p2 a owl:FunctionalProperty .
                ?p (rdfs:subPropertyOf|owl:equivalentProperty)+ ?p2 . }}}
            ?s ?p ?o . }}}
```

Listing 4: Requête SPARQL pour instancier l'ontologie sur la propriété FunctionalProperty

5 Conclusion

Dans cet article, nous avons présenté une étude à grande échelle concernant l'utilisation de la sémantique dans les graphes de connaissances des données liées ouvertes. Nous avons confirmé que l'utilisation de la sémantique est pour le moins sporadique. Malgré tout, certains graphes utilisent de nombreuses fonctionnalités de OWL 2. Nous proposons dans ce travail une ontologie permettant à l'éditeur de données de fournir aux utilisateurs du graphe des informations précieuses concernant l'utilisation de la sémantique dans le graphe de connaissances. Ainsi avisé, l'utilisateur pourra choisir selon ses besoins l'outil approprié à la réalisation de sa tâche. Nous prévoyons de proposer un outil permettant d'instancier automatiquement cette ontologie pour un graphe donné, quelle que soit sa sérialisation (fichier HDT ou turtle, point d'accès SPARQL, etc.). Nous prévoyons aussi de fournir les instances de cette ontologie pour les graphes proposés par le *LOD Cloud* et les fichiers HDT du *LOD Laundromat*.

Références

Alexander, K., R. Cyganiak, M. Hausenblas, et J. Zhao (2009). Describing linked datasets. In *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009.*

Beek, W., L. Rietveld, H. R. Bazoobandi, J. Wielemaker, et S. Schlobach (2014). LOD laundromat : a uniform way of publishing other people's dirty data. In *International Semantic Web Conference*, pp. 213–228. Springer.

d'Aquin, M., C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, et E. Motta (2007). Characterizing knowledge on the semantic web with watson.

Färber, M., F. Bartscherer, C. Menne, et A. Rettinger (2016). Linked Data quality of DBpedia, Freebase, OpenCyc, Wikidata, and Yago. *Semantic Web* (Preprint), 1–53.

Fernández, J. D., M. A. Martínez-Prieto, C. Gutiérrez, A. Polleres, et M. Arias (2013). Binary RDF representation for publication and exchange (HDT). *Web Semantics : Science, Services and Agents on the World Wide Web 19*, 22–41.

Hitzler, P. et F. van Harmelen (2010). A reasonable semantic web. *Semantic Web 1*, 39–44.

Hogan, A., A. Harth, A. Passant, S. Decker, et A. Polleres (2010). Weaving the pedantic web. *LDOW 628*.

Jain, P., P. Hitzler, P. Z. Yeh, K. Verma, et A. P. Sheth (2010). Linked data is merely more data. In *AAAI Spring Symposium : Linked Data Meets Artificial Intelligence*.

Martínez-Prieto, M. A., M. Arias, et J. D. Fernández (2012). Exchange and consumption of huge RDF data. In *The Semantic Web : Research and Applications*, pp. 437–452. Springer.

Polleres, A., A. Hogan, A. Harth, et S. Decker (2010). Can we ever catch up with the web ? *Semantic Web 1*, 45–52.

Weibel, S., J. A. Kunze, C. Lagoze, et M. Wolf (1998). Dublin core metadata for resource discovery. *RFC 2413*, 1–8.

Summary

The number and size of RDF knowledge graphs are constantly increasing. As a result, data processing for agents (automated or human) is becoming more and more difficult. If several tools can be used for a given task, but each depends to varying degrees on the semantics available in the knowledge graph, then it is important to have an overview before the graph to select the best tool for that task. We conducted a large-scale in-depth study to verify the presence of semantics in knowledge graphs currently published in the Linked Data. Although some knowledge graphs use OWL 2 semantics, many do not do so or only partially. We therefore propose an approach that, based on statistics, instantiates an ontology that facilitates the selection of the most suitable tool for a given task based on the use of OWL 2 semantics.