Évaluation de la qualité des sources du web de données pour la résolution d'entités nommées

Carmen Brando [1], **Nathalie Abadie** [2], **Francesca Frontini** [3]

1. *CRH UMR 8558 CNRS - EHESS*
   Paris, France
   carmen.brando@ehess.fr
2. *Univ. Paris-Est, LASTIG COGIT, IGN, ENSG*
   Saint-Mandé, France
   nathalie-f.abadie@ign.fr
3. *Praxiling UMR 5267 CNRS - UPVM3*
   Université Paul-Valéry Montpellier 3, France
   francesca.frontini@univ-montp3.fr

RÉSUMÉ. *Les applications d'édition numérique de textes mettent à profit les URI du web de données afin d'identifier les entités nommées mentionnés et encore pour accéder à des informations complémentaires sur ces entités. On appelle résolution d'entités nommées la tâche qui consiste à assigner automatiquement une référence choisie au sein d'une base de connaissances à une mention d'entité nommée préalablement étiquetée dans un texte. Cependant, les sources de données du web de données mises à contribution pour ce type d'applications peuvent présenter des problèmes de qualité ayant des conséquences néfastes sur les résultats obtenus. Dans cet article, nous présentons une étude empirique réalisée afin d'évaluer la qualité de jeux de données du web de données en tant que bases de connaissances potentielles pour une application de résolution d'entités nommées dans le contexte des humanités numériques. Pour ce faire, nous nous appuyons sur des mesures d'évaluation de la qualité des sources de données du web de données de l'état de l'art mises en œuvre du point de vue de l'adéquation des données à un besoin particulier. Nous testons ces mesures sur des sources de données de deux types : une source de données du web de données généraliste et d'autres portant sur des domaines plus spécifiques. L'objectif visé est de déterminer s'il est possible d'évaluer a priori laquelle de ces sources de données sera la plus à même de produire de bons résultats de résolution d'entités nommées dans le cas de textes littéraires en français.*

ABSTRACT. *More applications in the Digital Humanities rely on Linked Data for the semantic enrichment of digital collections by means of URI, typically for providing background information about authors, works of art and historical places, mentioned in these collections. In this sense, Named Entity Linking (NEL) is the task of automatically assigning the appropriate referent to*

a named-entity mention tagged in a text. Nevertheless, data sources of the Web of Data still experiences quality issues which are critical for NEL and many Digital Humanities applications. The present article hence proposes an empirical study to assess the quality of any Linked Data (LD) set meant to be used as Knowledge Base in graph-based NEL. Our methodology deals with state-of-art quality aspects from a fitness-for-use perspective. We perform experiments on two French heritage texts and choose to test two types of linking: on the one hand to a generalistic Linked Data source and on the other to domain-specific ones. The proposed study assesses to which degree the different Linked Data sources are better suited to be used as Knowledge Base for some NEL use case.

MOTS-CLÉS : *qualité des données, résolution d'entités nommées, web des données, humanités numériques.*

KEYWORDS: *data quality, named-entity linking, Linked Data, digital humanities.*

DOI:10.3166/ISI.21.5-6.31-54

1. Introduction

La disponibilité croissante de ressources publiées sur le web de données ouvre de nouvelles perspectives pour le développement d'applications. En particulier, de plus en plus d'applications d'édition numérique comme Recogito [1], Pundit [2] ou Reden Online [3] destinées à l'annotation semi-automatique de textes, mettent à profit des URI aussi bien pour identifier des entités nommées comme des écrivains, des œuvres d'art, des lieux historiques ou des organisations mentionnées dans des textes que pour référencer des ressources externes. L'attribution d'identifiants à des mentions d'entités nommées vise à permettre l'enrichissement sémantique de documents en leur adjoignant des connaissances supplémentaires et en favorisant la désambiguïsation des informations qu'ils renferment. Dans le cas où ces identifiants proviennent de ressources publiées sur le web de données, il devient en outre possible d'aller rechercher à la volée des informations supplémentaires en suivant les liens owl [4]:sameAs reliant les ressources à leurs équivalents sur le web de données. De plus, la structuration en graphe des ressources et la disponibilité des vocabulaires les décrivant facilitent le développement de solutions de visualisation, de recherche par facettes et d'accès aux contenus particulièrement utiles dans le domaine des humanités numériques.

La tâche dite de *résolution d'entités nommées* (que nous désignerons par l'acronyme NEL pour *Named Entity Linking*) consiste à affecter automatiquement la bonne référence, l'identifiant de la ressource correspondante dans une base de connaissances ou bien NIL (l'absence de valeur), à une entité nommée préalablement identifiée dans un texte et étiquetée. Les approches de résolution d'entités nommées non supervisées

1. http://recogito.pelagios.org
2. http://thepund.it
3. http://obvil-dev.paris-sorbonne.fr/reden/RedenOnline/site/input-tei.html
4. http://www.w3.org/2002/07/owl#

comme Han *et al*. (2011), Usbeck *et al*. (2014), Brando *et al*. (2015) et Besançon *et al*. (2016) s'articulent le plus souvent en deux étapes successives (Mihalcea, Csomai, 2007), dites de *sélection des candidats* et de *classement des candidats*. Chacune de ces étapes nécessite de disposer de bases de connaissances externes ; de plus en plus souvent, ce sont des sources de données du web de données qui sont mises à profit. Cependant, les bases de connaissances publiées sur le web de données ne présentent pas nécessairement toutes le même niveau de qualité au regard des applications auxquelles on les destine, et le choix de l'une ou l'autre des bases disponibles pour une application donnée aura donc des conséquences sur les résultats de cette dernière (Paulheim, Bizer, 2014).

Ainsi, dans le cadre des méthodes non supervisées pour la résolution d'entités nommées, l'application en humanités numériques est influencée par la spécificité du contexte. Ceci est caractérisé par une hétérogénéité externe, par rapport aux corpus contemporains et journalistiques normalement utilisés en NEL mais aussi interne, car les textes sont très éloignés les uns des autres du point de vue historique ainsi que du sujet traité. En effet, il peut s'agir d'un texte sur les romanciers ou poètes français ou bien les œuvres d'art contemporains et concerner des périodes bien déterminées comme le 19e siècle ou l'Antiquité et des zones géographiques particulières comme la France, l'Asie ou encore le monde entier. Dans ce contexte, la portée de la base de connaissances choisie aura une importance considérable sur les résultats de l'application. De la même façon, un manque d'exhaustivité dans la description des ressources de la base de connaissances pourra conduire à une baisse significative des résultats de l'application. Ainsi, l'absence d'étiquettes multilingues, de formes rejetées [5] comme les noms historiques, les abréviations, les surnoms ou encore de pseudonymes pourront nuire à l'étape de sélection des candidats. Dans les cas où l'étape de classement des candidats se fonde sur le graphe formé par l'ensemble des ressources candidates et les propriétés qui les relient, l'absence de liens entre ces ressources pourra également avoir un effet néfaste sur la bonne désambiguïsation des entités nommées. Enfin, l'absence de liens d'équivalence (du type owl:sameAs ou skos [6]:exactMatch, par exemple) avec d'autres ressources du web de données, privant l'application d'informations complémentaires, pourra également porter préjudice à son bon fonctionnement.

La plupart des travaux existants dans le domaine de l'évaluation de la qualité des données liées portent sur l'évaluation de la qualité des grandes bases de connaissances généralistes comme DBpedia ou Yago tandis que peu d'attention a été donnée à celle de sources de données plus spécialisées sur un domaine de connaissances en particulier (Erp *et al*., 2016). Contrairement aux grandes bases de connaissances généralistes, généralement construites à partir de données non structurées, les sources de données plus spécialisées ont généralement été produites par des institutions publiques (musées, bibliothèques, agences cartographiques, instituts de statistiques, etc.), en suivant

5. Une forme rejetée représente une appellation alternative de personnes, des noms géographiques ou des œuvres, etc. dans une notice d'autorité, à l'inverse, la forme retenue est l'appellation utilisée de préférence.
6. http://www.w3.org/2004/02/skos/core#

des chaînes d'acquisition, de structuration, de mise à jour propres, pour des usages souvent différents de ceux des applications qui se développent actuellement autour des données du web. Elles présentent donc des propriétés différentes en termes de qualité et leur réutilisation optimale nécessite qu'on y prête une attention particulière.

Dans cet article, nous nous employons donc à évaluer la qualité des bases de connaissances publiées sur le web de données dans le cadre d'applications de résolution d'entités nommées dans le domaine des humanités numériques. Notre objectif est de favoriser le choix des ressources les plus pertinentes en amont du processus de résolution d'entités nommées afin d'améliorer les résultats d'applications automatiques d'annotation et d'enrichissement sémantiques d'éditions numériques d'œuvres littéraires françaises. Nous nous attacherons donc tout d'abord à dresser un état de l'art des travaux liés à l'évaluation de la qualité de sources de données du web de données en vue de leur utilisation pour des applications particulières. Puis, nous décrirons le fonctionnement global d'un processus de résolution d'entités nommées non supervisé et les mesures les plus fréquemment utilisées afin d'en évaluer les résultats. Nous restreignons notre analyse aux approches non supervisées car les approches supervisées restent peu répandues pour les applications liées aux humanités numériques, faute de textes préalablement annotés en nombre suffisant. Ensuite, nous tâcherons de dégager de cette description les aspects de la qualité des sources de données du web de données à privilégier pour favoriser le bon fonctionnement du processus de résolution d'entités nommées sur un texte particulier. Nous proposerons un ensemble de mesures d'évaluation de la qualité de sources de données du web de données issues de la littérature, choisies et adaptées afin de refléter quantitativement les avantages et inconvénients de chaque base de connaissance. Afin de nous assurer de la pertinence de ces mesures, nous confronterons enfin les résultats de ces mesures avec les résultats réels obtenus par une application de résolution d'entités nommées non supervisée sur deux textes distincts.

2. Mesures pour l'évaluation des résultats des applications de résolution d'entités nommées

En règle générale, les applications de résolution d'entités nommées non supervisées se déroulent en deux phases (Mihalcea, Csomai, 2007; Hachey *et al*., 2011). La première étape, dite de *sélection des candidats*, inclut les tâches d'extraction des données et de recherche des candidats. La seconde, dite de *classement des candidats*, consiste à identifier parmi les candidats présélectionnés les plus susceptibles de correspondre effectivement à l'entité nommée mentionnée dans le texte. Dans la suite, nous décrivons plus en détail le fonctionnement de chacune de ces étapes et les mesures généralement utilisées afin d'évaluer leurs résultats respectifs par comparaison avec un texte de référence annoté manuellement (souvent appelé *gold standard*).

2.1. Évaluation des résultats de la phase de sélection des candidats

La première étape d'un processus de résolution d'entités nommées non supervisé vise tout d'abord à construire un dictionnaire (voire un index) de références potentielles définies à partir des ressources d'une base de connaissances externe. Cette tâche est réalisée en amont du processus global et le dictionnaire produit peut par la suite être réutilisé dans le cadre de l'annotation d'autres textes mentionnant des entités relevant du même domaine applicatif. Ce dictionnaire est généralement produit à partir de données issues de fichiers de données téléchargeables ou bien obtenues par requêtes sur un point d'accès SPARQL. Puis, pour chaque entité nommée mentionnée dans le texte et préalablement étiquetée, des entrées candidates sont sélectionnées dans le dictionnaire, le plus souvent par comparaison de chaînes de caractères, à l'aide d'heuristiques de similarité ou par un mécanisme d'expansion de requêtes (ici, la mention) utilisé en Extraction d'Information. On parle alors d'ensembles de candidats. Les résultats de cette étape de sélection des candidats sont généralement évalués à l'aide de mesures comme celles décrites par Hachey *et al*. (2011) et complétées par Brando *et al*. (2016).

La *cardinalité moyenne des ensembles de candidats (CardM)* est calculée à partir des ensembles des candidats extraits du dictionnaire pour chaque mention d'entité nommée étiquetée. Cet indicateur reflète la capacité du dictionnaire à fournir des candidats pour l'étape de classement à suivre: une valeur trop faible indique un dictionnaire incomplet par rapport à l'univers du discours du texte, tandis qu'une valeur trop élevée est le signe d'une surabondance d'entrées dotées d'étiquettes homonymes dans le dictionnaire qui risquent de nuire à l'étape de désambiguïsation à suivre. Ainsi, l'amélioration des performances de l'étape de sélection des candidats requiert d'une part, de retrouver le candidat pertinent au sein du dictionnaire lorsque celui-ci est bien référencé dans la base de connaissances externe, et d'autre part, d'éviter la sélection inutile de candidats erronés afin de réduire l'effort de désambiguïsation à fournir à l'étape suivante.

Ce premier aspect est plus précisément évalué en calculant le *rappel des candidats (RapCand)*. Il s'agit de la proportion des ensembles de candidats non vides et contenant la bonne référence renvoyés par l'étape de sélection des candidats pour chaque mention d'entité nommée étiquetée par rapport au nombre total de mentions pour lesquelles il existe effectivement une référence pertinente dans la base de connaissances. La capacité du processus de sélection à retrouver des candidats pertinents lorsque les références correctes sont effectivement disponibles dans la base de connaissances est également évaluée par la mesure dite de *précision des références de type NIL (PrecN)*. Celle-ci calcule la proportion des ensembles de candidats vides renvoyés par le processus de sélection lorsque la bonne référence ne figure effectivement pas dans la base de connaissances par rapport au nombre total d'ensembles de candidats vides renvoyés.

La *précision des candidats (PrecCand)* calcule la proportion des ensembles de candidats non vides et contenant la bonne référence renvoyés par l'étape de sélection des candidats par rapport au nombre total d'ensembles de candidats non vides qu'elle

renvoie. En d'autres termes, un score de précision des candidats élevé indique que l'étape de sélection procure majoritairement des candidats pertinents et peu de bruit. Enfin, le *rappel des références de type NIL (RapN)* évalue la proportion des ensembles de candidats vides renvoyés par l'étape de sélection lorsque la base de connaissances ne contient effectivement pas de référence pertinente pour la mention recherchée par rapport au nombre total de mentions pour lesquelles il n'existe pas de référence pertinente dans la base de connaissances.

2.2. Évaluation des résultats de la phase de classement des candidats

La seconde étape d'un processus de résolution d'entités nommées non supervisé consiste à choisir les candidats les plus pertinents pour chaque mention d'entité nommée étiquetée. Dans le cas des approches non supervisées, les approches fondées sur les graphes de données RDF suscitent un intérêt croissant (Han *et al*., 2011; Usbeck *et al*., 2014; Brando *et al*., 2015; Besançon *et al*., 2016). Elles reposent sur les triplets constituant la base de connaissances et sont fondées sur les notions de centralité de graphe (Brando *et al*., 2015), de sous-graphe le plus dense (Moro *et al*., 2014), de parcours aléatoires du graphe (Gruetze *et al*., 2016) ou encore de représentation des relations sous la forme d'un vecteur de contexte relationnel Besançon *et al*. (2016). Les scores de pertinence des candidats sont généralement calculés à l'aide d'algorithmes de classement de pages web, de comparaisons de vecteurs de contexte relationnel ou de calcul du degré de centralité appliqués sur le graphe défini à partir des données RDF de la base de connaissances et dans lequel les nœuds correspondent aux ressources candidates et les arcs aux propriétés qui les relient au sein de cette même base de connaissances. La figure 1 illustre le résultat du calcul des scores de centralité d'un algorithme de résolution d'entités nommées tel que REDEN à partir d'un extrait d'un texte (voir Brando *et al*. (2016) pour plus de détails) dont six mentions d'auteurs sont préalablement étiquetées chacune possédant au moins un candidat dans la base de connaissances. Nous observons que les nœuds *FrenchPoets*[7] et *SymbolistPoets*[8] sont ceux qui influencent le choix final. Autrement dit, le candidat choisi par mention possède la valeur de score de centralité la plus élevée sur l'ensemble de ses concurrents. En effet, le texte de l'exemple fait référence à des poètes français. Parmi ceux cités, Charles Baudelaire et Paul Verlaine comptent parmi les figures les plus remarquables de leur époque. Ces deux poètes appartenaient au mouvement du Symbolisme, tout comme Jules Laforgue. Dans d'autres cas, il est possible de s'appuyer sur des propriétés telles que http://dbpedia.org/ontology/influencedBy ou http://dbpedia.org/ontology/influenced qui peuvent parfois jouer un rôle important dans le choix du bon candidat.

En outre, les résultats de cette étape de classement des candidats dépendent de ceux de l'étape de sélection des candidats. La mesure d'*exactitude de désambiguï-*

7. http://dbpedia.org/class/yago/FrenchPoets
8. http://dbpedia.org/class/yago/SymbolistPoets

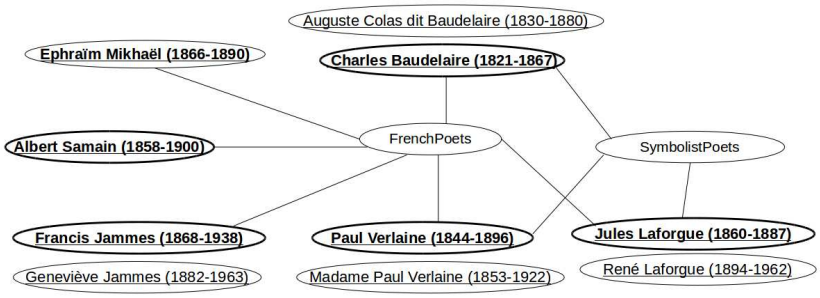

Figure 1. Extrait des URI choisis (en gras) pour les neuf candidats (soulignés); tous les arcs sans étiquette représentent des liens rdf: type; cette figure est extraite de Brando et al. *(2016)*

sation (ExactD) vise à fournir une estimation des résultats de l'étape de classement indépendante de ceux de l'étape de sélection. Ainsi, il s'agit de la proportion des références correctement assignées lorsque l'ensemble des candidats sélectionnés à la première étape pour chaque mention contient effectivement la bonne référence.

Enfin les résultats du processus global de résolution d'entités nommées peuvent également être synthétisés à l'aide de mesures comme *l'exactitude globale (ExactG)*. Il s'agit de la proportion de références correctement assignées pour chaque mention d'entité nommée, lorsque ces dernières disposent effectivement d'une référence pertinente dans la base de connaissances. En outre, la proportion des références de type NIL assignées à bon escient peut venir compléter cette mesure.

3. Influence de la qualité des sources du web de données sur les applications de résolution d'entités nommées

Les résultats d'une application de résolution d'entités nommées non supervisée dépendent donc non seulement des algorithmes mis en œuvre, mais également beaucoup des bases de connaissances utilisées pour alimenter le dictionnaire des candidats et pour générer le graphe support de l'algorithme de classement des résultats. Dans ce contexte, il semble donc utile d'identifier et d'évaluer les éléments de qualité que doivent présenter les bases de connaissances afin de guider le choix d'une (ou plusieurs) base(s) de connaissances qui garantissent un fonctionnement optimal à chacune des étapes du processus.

Au cours des dernières années, la communauté de recherche en web sémantique s'est plus particulièrement intéressée à l'évaluation de la qualité des données publiées sur le web. La plupart des travaux engagés dans ce sens visent à établir et améliorer la qualité des sources de données liées à l'aide d'approches automatiques ou semi-automatiques fondées sur le profilage de données (Jentzsch *et al*., 2015), l'analyse de leur distributions statistiques (Paulheim, Bizer, 2014), les réseaux bayésiens

(Ruckhaus *et al*., 2014) ou encore la définition de formalismes comme *RDF Description Language* pour la description de contraintes d'intégrité au niveau des instances (Schmidt, Lausen, 2013). Chacune de ces propositions traite des questions de redondance au sein des sources de données et de détection d'incohérences entre les instances et les vocabulaires qui les décrivent.

Zaveri *et al*. (2015) présentent un état de l'art extrêmement détaillé des mesures proposées dans la littérature pour évaluer la qualité des jeux de données liées en termes de facilité d'accès aux données, de qualité interne des données, de confiance, d'actualité et de gestion des versions et enfin, d'adéquation aux usages. Les aspects de la qualité liés à la qualité interne des sources du web de données, comme leur correction ou encore leur cohérence, pouvant avoir une influence sur les résultats de toutes les applications mettant en œuvre ces sources de données, sont très largement traités dans la littérature et une synthèse détaillée de leurs différents aspects est proposée dans Zaveri *et al*. (2015). C'est pourquoi nous nous attacherons ici à identifier et évaluer en priorité les aspects de la qualité des sources du web de données portant sur l'adéquation de ces dernières à des applications de résolution d'entités nommées non supervisées.

3.1. Influence de la qualité des sources de données du web de données sur l'étape de sélection des candidats

L'étape de sélection des candidats repose sur la détection automatique de relations de correspondance entre les mentions d'entités nommées du texte analysé et les entrées du dictionnaire construit à partir des données d'une base de connaissances.

3.1.1. Portée des données

La recherche de correspondances entre mentions d'entités nommées du texte et entrées du dictionnaire suppose tout d'abord un bon recouvrement thématique entre l'univers du discours du texte et les données sélectionnées dans la base de connaissances et importées dans le dictionnaire. En effet, afin de se doter d'un dictionnaire permettant la résolution du plus grand nombre d'entités nommées mentionnées dans le texte, il convient de sélectionner dans la base de connaissances le plus possible de ressources. Il convient également d'exclure *à priori* celles qui ne sont pas pertinentes afin de réduire, autant que faire se peut, les risques de sélection de candidats homonymes non pertinents lors de la première étape du processus de résolution d'entités nommées. Cette question de la portée des données utilisées en entrée de la création du dictionnaire correspond à deux éléments de qualité des données du web de données que Zaveri *et al*. (2015) nomment «pertinence» (*relevancy*) et «quantité de données» (*amount-of-data*). En effet, le filtrage efficace des données en entrée de la création du dictionnaire nécessite de disposer de données instanciant les propriétés et de concepts appropriés pour effectuer la sélection souhaitée. Par exemple, si le texte à traiter porte sur des écrivains français du 19e siècle et leurs influences littéraires, il sera nécessaire de disposer de la date de naissance des ressources de type «Personne» que l'on souhaite inclure dans le dictionnaire afin d'éliminer les personnes qui s'avèrent nées après l'année 1900. Sans cette information, tout filtrage temporel sera impossible et

le dictionnaire devra comporter l'ensemble des ressources de type «Personne» de la base de connaissances, y compris celles dont on sait *a priori* qu'elles ne concernent pas l'univers du discours du texte. Dans le même ordre d'idées, le manque de typage des ressources constitue un frein à la sélection préalable des ressources les plus pertinentes pour une application, voire peut engendrer des erreurs. Par exemple, la ville de Berlin n'est à ce jour considérée dans le chapitre francophone de DBpedia ni comme un lieu de peuplement (PopulatedPlace [9]) ni comme une entité spatiale (SpatialThing [10]) ce qui interdit son insertion dans un dictionnaire constitué de ressources de l'un ou l'autre de ces types. De la même façon des ressources typées uniquement à l'aide des concepts situés dans les niveaux les plus élevés des hiérarchies de concepts des vocabulaires utilisés ne pourront être importés sans ajouter du bruit dans le dictionnaire. Enfin, certains concepts définis dans le vocabulaire de la base de connaissances peuvent n'avoir que peu ou pas d'instances ce qui les rend inexploitables pour le filtrage des données du dictionnaire. Par exemple, la classe «Personnage de fiction» (FictionalCharacter [11]) de l'ontologie DBpedia ne peut pas être mise à profit faute d'un nombre d'instances suffisant, alors même que l'on rencontre fréquemment des mentions à des personnages imaginaires dans des textes.

3.1.2. Exhaustivité de population

L'exhaustivité de population d'une base de connaissances est un aspect de sa qualité proche de la notion de portée des données. Il s'agit du pourcentage d'entités du monde réel figurant dans la base de connaissances (voir la définition de la mesure dite d'«exhaustivité de population», *population completeness*, dans Zaveri *et al*. (2015)). Par exemple, dans le cas de textes citant des écrivains français du 19e siècle, les mentions à des écrivains célèbres sont relativement fréquentes. En revanche, ceux-ci ne figurent pas nécessairement tous dans les sources du web de données que l'on pourrait vouloir exploiter.

3.1.3. Richesse des dénominations

L'établissement des correspondances entre mentions du texte et entrées du dictionnaire est le plus souvent réalisé sur la base de mesures de similarité de chaînes de caractères. Les approches syntaxiques d'évaluation de la similarité de chaînes de caractères considèrent les paires de chaînes de caractères à comparer comme des séquences de lettres et de symboles graphiques, dont elles cherchent à déterminer si celles-ci désignent ou non la même chose en évaluant l'importance de leurs portions communes par rapport à leurs portions différentes. Shvaiko, Euzenat (2005), Ferraram *et al*. (2013) ou encore Cheatham, Hitzler (2013) présentent les mesures de similarité de chaînes de caractères les plus fréquemment utilisées pour la mise en correspondance d'étiquettes hétérogènes de ressources du web de données. Cependant, lorsqu'une même entité nommée peut être désignée par des expressions dont les diffé-

9. http://dbpedia.org/ontology/PopulatedPlace
10. http://www.w3.org/2003/01/geo/wgs84_pos#SpatialThing
11. http://dbpedia.org/ontology/FictionalCharacter

rences dépassent la simple variation orthographique, l'omission, l'abréviation ou encore la permutation de termes, ces mesures syntaxiques s'avèrent inefficaces. Des opérations de prétraitement sémantiques, comme la substitution des chaînes de caractères à comparer par des synonymes ou des traductions sont alors nécessaires (Cheatham, Hitzler, 2013). De telles opérations nécessitent de recourir à des ressources externes fournissant les connaissances utiles sur les synonymes ou les traductions possibles pour les chaînes de caractères à comparer. Dans le cas des applications de résolution d'entités nommées, ces synonymes ou ces traductions proviennent le plus souvent de la base de connaissances utilisée et leur mise à profit est réalisée en les intégrant directement au dictionnaire de candidats. L'abondance de dénominations alternatives pour les ressources de la base de connaissances constitue donc un avantage pour permettre la construction d'un dictionnaire favorisant la sélection d'entrées pertinentes pour chaque mention du texte. Cet aspect de la qualité des sources de données du web de données correspond à la fois à la notion de «versatilité» (*versatility*) d'un jeu de données (Zaveri *et al.*, 2015) qui concerne plus précisément la disponibilité d'étiquettes multilingues et à celle d'«exhaustivité des propriétés» (*property completeness*) qui désigne le rapport entre le nombre de valeurs assignées à une propriété et le nombre total de valeurs que celle-ci pourrait prendre. Dans le cas d'une application de résolution d'entités nommées issues de textes littéraires en français, la disponibilité dans la base de connaissances utilisée de formes rejetées, d'orthographes alternatives, ou encore de surnoms ou de pseudonymes en français constitue donc un atout pour la phase de sélection des candidats en favorisant la découverte des entrées pertinentes au sein du dictionnaire.

3.2. Influence de la qualité des sources de données du web de données sur l'étape de classement des candidats

L'étape de classement des candidats vise à ordonner les candidats sélectionnés lors de l'étape précédente en fonction de leur pertinence en tant que référence pour la mention d'entité nommée à annoter. Dans le cas des approches non supervisées, ce classement est généralement réalisé à l'aide d'algorithmes fondés sur le graphe formé par les ressources candidates (que nous désignerons par la suite comme des «nœuds candidats»), les ressources du vocabulaire utilisé pour typer les ressources candidates (qui nous appellerons «nœuds descriptifs») et les propriétés qui les lient (que nous considérerons comme les arcs de ce graphe). La présence d'arcs ou encore de nœuds descriptifs communs à plusieurs nœuds candidats seront donc susceptibles de favoriser le calcul de scores de classement plus pertinents. En d'autres termes, plus un graphe de ressources candidates comporte de connexions entre nœuds candidats et plus il sera susceptible d'assigner les références correctes aux mentions d'entités nommées à annoter. De plus, il est possible d'attribuer des poids à certains arcs afin de leur donner plus d'importance. Il est donc particulièrement important à cette étape de disposer de sources de données externes dotées de nombreux arcs et nœuds descriptifs. Ceux-ci pouvant éventuellement provenir de différentes sources de données, la présence de liens d'équivalence comme owl:sameAs ou skos:exactMatch au sein de la principale

base de connaissances choisie constituera également un atout pour la réussite de cette étape du processus de résolution d'entités nommées.

3.2.1. Granularité des concepts définis dans les vocabulaires décrivant les données

La granularité ou encore le niveau d'exhaustivité des concepts recensés dans les vocabulaires utilisés pour typer les ressources au sein d'une base de connaissances fournit des informations sur le contexte des ressources candidates et contribue donc à leur classement optimal. En effet, les candidats sont tout d'abord sélectionnés à l'aide de critères très généraux. La découverte, pour certains candidats de différentes mentions, de propriétés communes plus spécifiques, comme le fait d'appartenir à un même concept très spécialisé, constitue un bon indice concernant leur pertinence et permet de leur attribuer un bon classement. Dans le même ordre d'idées, Paulheim, Bizer (2014) ont étudié la question de la profondeur des concepts utilisés pour typer les données au sein de la hiérarchie de concepts du vocabulaire dans lequel ils sont définis. Dans Jentzsch *et al*. (2015), les auteurs analysent la répartition des ressources entre les différents niveaux de la hiérarchie des classes des vocabulaires qui les décrivent, pour les sources de données DBpedia et LinkedGeoData. Il soulignent la faible utilisation des classes des niveaux les plus bas des hiérarchies de concepts pour typer les ressources au sein de ces deux sources de données. Dans le cas des sources de données destinées à des applications de résolutions d'entités nommées non supervisées, il ne s'agit pas seulement de déterminer le nombre de niveaux dans la hiérarchie de concepts du vocabulaire sous-jacent des données, mais également d'évaluer à quel point les concepts couramment utilisés pour typer les ressources sont spécialisés ou non. En règle générale, les concepts les plus haut placés dans la hiérarchie comme la notion de «Personne», comptent un grand nombre de ressources. À l'inverse, les concepts les plus spécialisés regroupent beaucoup moins de ressources mais celles-ci partagent des propriétés bien plus spécifiques. C'est le cas par exemple des instances de la catégorie Yago «Poète symboliste» qui correspond au sous-ensemble des écrivains appartenant au mouvement symboliste. L'utilisation de ressources typées à l'aide de concepts très spécialisés constitue donc un atout pour le bon fonctionnement des algorithmes de classement dans la mesure où ceux-ci prouvent que les candidats comparés partagent des propriétés communes particulièrement discriminantes. Il est donc important de s'assurer, lors du choix d'une base de connaissances pour la résolution d'entités nommées, que celle-ci est décrite par un vocabulaire suffisamment détaillé et que les concepts définis par ce vocabulaire sont bien effectivement utilisés pour typer les ressources.

3.2.2. Présence effective de propriétés entre ressources du même type

De la même façon, le bon fonctionnement des algorithmes de classement fondés sur les graphes nécessite de disposer de suffisamment d'arcs entre les nœuds constitués par les ressources candidates. Ceci suppose donc de disposer, au sein de la base de connaissances utilisée pour créer le graphe des candidats, de propriétés entre ressources du même type et susceptibles d'être sélectionnées comme candidates. Dans le cas de sources du web de données il s'agit donc de privilégier les sources de données

au sein desquelles on dispose de nombreuses propriétés instanciées entre ressources du même type. Si nous reprenons l'exemple des textes sur les écrivains français du 19e siècle, l'existence de propriétés entre ressources de type «Écrivain» comme c'est le cas dans DBpedia avec la relation http://dbpedia.org/ontology/influencedBy, constituera un atout pour le classement des candidats. De façon plus générale, la présence de propriétés non seulement entre ressources typées par des concepts relevant de différents niveaux hiérarchiques, mais également entre ressources du même type peut être mise à profit dans de nombreuses applications de désambiguïsation fondées sur les graphes, comme par exemple les applications de désambiguïsation de termes (Sinha, Mihalcea, 2007). En sémantique computationnelle, notamment dans les études sur la polysémie, les questions liées à la représentation et à la manipulation de l'information sont primordiales. De nombreuses théories sur les lexiques, comme celle décrite par Pustejovsky (1991), mettent l'accent sur les relations transverses entre concepts pour permettre l'interprétation de la sémantique des termes qui les désignent.

3.2.3. Exhaustivité des liens d'équivalence entre ressources

Enfin, l'exhaustivité des liens d'équivalence entre ressources est un aspect de la qualité interne des sources du web de données bien connu et traité dans de nombreux travaux décrits en détail dans l'état de l'art de Zaveri *et al*. (2015). Il s'agit d'évaluer à quel point les ressources d'un jeu de données sont interconnectées avec des ressources équivalentes issues de jeux de données externes. Dans le cas des sources du web de données, cela revient à compter le nombre de liens de type owl:sameAs (ou skos:exactMatch) entre ressources. Dans le cas d'une application de résolution d'entités nommées, la présence de tels liens dans la base de connaissances sur laquelle repose l'application peut être mise à profit afin d'aller rechercher à la volée des informations complémentaires dans d'autres sources de données. Elle constitue donc un atout pour améliorer les résultats de l'ensemble du processus.

4. Mesures d'évaluation de la qualité des sources du web de données pour des applications de résolution d'entités nommées

Dans la section précédente, nous avons dressé l'inventaire des éléments de qualité à prendre en considération lors du choix d'une base de connaissances destinée à alimenter une application de résolution d'entités nommées. Nous nous attachons ici à proposer pour chacun de ces aspects de la qualité des sources du web de données, une ou plusieurs mesures d'évaluation directement choisies parmi ou dérivées des mesures de l'état de l'art afin de refléter quantitativement le potentiel de chaque base de connaissance. Enfin, nous commentons, pour chacune de ces mesures, les implications probables entre les valeurs obtenues par une source de données du web de données et les résultats d'une application de résolution d'entités nommées mettant en œuvre cette source de données.

4.1. Mesures d'évaluation de la qualité des sources de données du web de données pour l'étape de sélection des candidats

L'amélioration des résultats de la phase de sélection des candidats d'un processus de résolution d'entités nommées suppose de disposer de sources de connaissances externes qui couvrent le plus précisément possible l'univers du discours du texte, qui soient exhaustives, et renferment le plus possible de dénominations alternatives afin de favoriser l'identification des ressources candidates les plus pertinentes.

L'estimation de la capacité d'une source de données à fournir les données nécessaires pour une application de résolution d'entités nommées particulière et exclusivement ces données, est une tâche relativement complexe. Elle correspond aux éléments de qualité dits de «pertinence» (*relevancy*) et de «portée» (*coverage*), décrit dans Zaveri *et al*. (2015). L'évaluation de la pertinence, fondée sur l'identification des sources de données pouvant être mises en œuvre pour une application et réellement utiles, est relativement complexe à réaliser de façon automatique. Zaveri *et al*. (2015) relèvent la subjectivité des rares mesures proposées dans la littérature pour évaluer la pertinence des sources du web de données pour une application précise. C'est pourquoi cette tâche est pour l'instant laissée à l'expert qui définit les critères de filtrage de la source de données. Cet aspect de pertinence est donc pour l'instant apprécié de façon plutôt qualitative, à travers les possibilités de filtrage des ressources selon leur thématique et leurs propriétés spatiales et temporelles offertes par la source de données. Les critères de filtrage définis par l'expert demeurent constants pour l'ensemble des calculs des autres mesures d'évaluation de la qualité de la source de données, afin de disposer de valeurs vraiment caractéristiques des données destinées à notre application de résolution d'entités nommées. En particulier, ces critères sont conservés pour évaluer la couverture du domaine d'intérêt fournie par la source de données, que nous appelons ici *portée des données*. Celle-ci est estimée à travers le nombre de ressources vérifiant les critères de sélection choisis par l'expert pour sélectionner les données correspondant *a priori* à l'univers du discours du texte à traiter. Plus formellement, la portée d'une source de données du web de données $Port_{LD}$ correspond au nombre de ressources de l'ensemble E_{LD} qui vérifient les critères de filtrage définis par l'expert et exprimés en termes de propriétés P et de valeurs assignées à ces propriétés V au sein de la source de données, c'est-à-dire,

$$Port_{LD} = | \{ E : E \; in \; E_{LD} \wedge (E, P, V) \} |$$

Cette mesure spécialise celle présentée dans Zaveri *et al*. (2015) qui dénombre les ressources d'une source de données sans filtrage préalable. En revanche, contrairement aux mesures proposées dans Zaveri *et al*. (2015), nous ne prenons pas en compte ici le nombre de propriétés disponibles au sein de la source de données.

Le choix de la source de données utilisée pour peupler le dictionnaire a des conséquences très importantes sur les résultats de la première étape du processus de résolution d'entités nommées. En effet, lorsque la portée du jeu de données utilisé pour alimenter le dictionnaire correspond parfaitement à l'univers du discours du texte, le

dictionnaire est très susceptible de contenir surtout des entrées pertinentes et très peu de bruit, ce qui permet d'espérer un bon score de rappel des références de type NIL. De plus, un jeu de données de portée trop large conduira à un dictionnaire doté d'entrées *a priori* non pertinentes mais potentiellement homonymes des mentions du texte ce qui pourrait engendrer un mauvais score de rappel des références de type NIL mais un bon score de précision des références de type NIL. A l'inverse, un jeu de données de portée trop restreinte aura un score d'exhaustivité de population potentiellement plus faible et conduira à plus de résultats de type NIL.

Pour évaluer l'exhaustivité d'une source de données, la métrique la plus couramment préconisée est celle nommée «exhaustivité de population» (*population completeness*) par Zaveri *et al*. (2015) et décrite comme la proportion des ressources disponibles dans la source de données par rapport au nombre total d'entités du monde réel portant sur un thème donné. Or l'évaluation d'un tel ratio nécessite de disposer de connaissances sur le nombre d'entités du monde réel correspondant à l'univers du discours du texte, ce qui est le plus souvent impossible. C'est pourquoi nous proposons de substituer à cette approche une mesure d'évaluation relative permettant d'appréhender l'exhaustivité de la source de données par rapport aux besoins d'annotation du texte à traiter. Ainsi, si nous considérons l'ensemble des mentions étiquetées du texte E_{text}, et l'ensemble des ressources d'une source de données du web de données vérifiant les critères de filtrage initiaux E_{Port}, l'exhaustivité de population PC_{LD} est définie comme la cardinalité de l'ensemble résultant de l'intersection des ressources vérifiant les critères de filtrage initiaux et de celles utilisées pour annoter les mentions d'entités dans une annotation de référence du texte :

$$PC_{LD} = \mid E_{text} \cap E_{Port} \mid$$

Ainsi, on pourra s'attendre à ce qu'une base de connaissances dotée d'une score d'exhaustivité de population élevé suscite des scores élevés de précision des candidats. A l'inverse, un score faible pourra être interprété comme annonçant un score de précision des références de type NIL élevé. En effet, comme les références de type NIL seront très fréquentes dans les annotations de référence, les ensembles de candidats vides renvoyés par le système seront plus susceptibles d'être pertinents.

La richesse d'une base de connaissances en termes de dénominations alternatives (LR) est évaluée en calculant le nombre moyen de valeurs assignées aux propriétés de type rdfs:label, skos:prefLabel ou encore skos:altLabel associées aux ressources disponibles dans la source de données et vérifiant les critères de filtrage initiaux. Nous distinguons les dénominations exprimées dans la langue de référence du texte à traiter de celles exprimées dans d'autres langues, *a priori* moins pertinentes pour une application de résolution d'entités nommées sur ce texte. Nous proposons ici encore une mesure différente de celle dite d'«exhaustivité de propriété» (*property completeness*) décrite par Zaveri *et al*. (2015) car cette dernière nécessite, comme celle d'exhaustivité de population, de disposer de connaissances sur le nombre total de valeurs possibles pour les propriétés qui nous intéressent dont il est difficile de disposer. Comme le score de rappel des candidats est très dépendant de la quantité et de la variété des dénomina-

tions alternatives disponibles dans le dictionnaire, les bases de connaissances avec des scores élevés pour ce critère de qualité seront plus susceptibles de produire un score élevé de rappel des candidats. De la même façon, pour les bases de connaissances dotées de scores faibles pour ce critère de richesse des dénominations alternatives, on pourra anticiper de nombreux échecs lors de la phase de sélection des candidats à partir du dictionnaire en raison d'un manque d'étiquettes différentes générées pour les entrées du dictionnaire. C'est pourquoi le score de précision des références de type NIL sera probablement assez bas dans ce cas de figure.

4.2. Mesures d'évaluation de la qualité des sources du web de données pour l'étape de classement des candidats

Dans les approches de résolution d'entités nommées non supervisées, le classement des candidats est généralement réalisé à l'aide d'algorithmes mettant à profit le graphe constitué par les ressources candidates et leurs propriétés. Pour choisir une base de connaissances à même de produire de bonnes performances à cette étape, nous proposons d'en évaluer la qualité à l'aide des mesures suivantes.

L'évaluation quantitative de la granularité des concepts définis dans les vocabulaires qui décrivent les données nécessite deux mesures. La première compte le nombre de concepts spécialisant le concepts principal utilisé dans les critères de filtrages initiaux pour choisir les données en entrée du processus de résolution d'entités nommées. Il s'agit donc du nombre de concepts instanciés par les ressources vérifiant les critères de filtrage initiaux et spécialisant le concept utilisé comme critère pour ce même filtrage. Plus formellement, la granularité d'une source de données du web de données $G1_{LD}$ correspond à la cardinalité de l'ensemble des concepts C instanciés par les ressources vérifiant les critères de filtrage initiaux E_{Port} *via* une propriété de typage comme rdf:type ou encore dcterms:subject et spécialisant le concept principal utilisé pour filtrer les données,

$$G1_{LD} = | \{ C : (E_{Port} \text{ a } C) \} |$$

La seconde mesure estime dans quelle mesure les ressources vérifiant les critères de filtrage initiaux instancient ou non ces concepts plus spécialisés. Elle peut donc être traduite comme la proportion des ressources vérifiant les critères de filtrage initiaux E_{Port} qui instancient l'un des concepts plus spécialisés identifiés dans le jeu de données.

$$G2_{LD} = | \{ E_{Port} : (E_{Port} \text{ a } C) \} | \div | E_{Port} |$$

La présence de propriétés entre ressources du même type est quantifiée par le nombre moyen de propriétés instanciées entre ressources vérifiant les critères de filtrage initiaux, pour chaque ressource de ce type. Ainsi, la mesure de la présence de propriétés entre ressources du même type dans une source de données du web de données PR_{LD} correspondant au nombre total de triplets comportant deux ressources

différentes vérifiant les critères de filtrage initiaux Ei_{Port} qui instancient une propriété Op, divisé par le nombre total de ressources vérifiant les critères de filtrage initiaux $|E_{Port}|$.

$$PR_{LD} = | \{ (E1_{Port}\ Op\ E2_{Port}) : E1_{Port} <> E2_{Port} \} | \div | E_{Port} |$$

De la même façon, l'exhaustivité des liens d'équivalence entre ressources est quantifiée par le nombre moyen de propriétés du type owl:sameAs, du type owl:sameAs ou skos:exactMatch, avec des ressources d'autres sources de données définies pour chaque ressource vérifiant les critères de filtrage initiaux. Cette mesure est relativement intuitive et fournit une valeur simple à interpréter. Elle est relativement proche de la mesure dite d'«exhaustivité de liage» (*interlinking completeness*) présentée par Zaveri *et al*. (2015) qui évalue le pourcentage des ressources d'un jeu de données qui possèdent des liens d'équivalence. Le calcul de l'exhaustivité des relations d'équivalence d'une source de données EE_{LD} revient à diviser le nombre des triplets ayant pour sujet une ressource vérifiant les critères de filtrage initiaux E_{Port}, pour prédicat une relation d'équivalence Ip, et pour objet une ressource d'une source de données externe E_{Ext}, par le nombre total de ressources vérifiant les critères de filtrage initiaux $|E_{Port}|$

$$EE_{LD} = | (E_{Port}\ Ip\ E_{Ext}) | \div | E_{Port} |$$

Les applications de résolution d'entités nommées non supervisées mettant généralement en œuvre des algorithmes fondés sur le graphe formé par les ressources candidates et leurs propriétés pour classer ces ressources, la présence d'arcs et de nœuds descriptifs entre candidats devrait influencer positivement le score d'exactitude de désambiguïsation. En d'autres termes, un graphe doté de nombreuses connexions entre ses nœuds sera plus susceptible de permettre l'identification des références correctes pour les mentions d'entités du texte. Il est donc primordial d'utiliser pour cette étape une base de connaissances présentant des scores élevés de granularité des concepts et de présence de relations entre ressources du même type. De façon plus générale, disposer de liens d'équivalence avec d'autres ressources permet d'en extraire des informations complémentaires lorsque la base de connaissance choisie ne remplit pas tous les critères de qualité désirés, en particulier ceux concernant la présence effective de propriétés entre ressources et la richesse des dénominations alternatives. Il est donc également souhaitable d'obtenir un score élevé pour la mesure de l'exhaustivité des relations d'équivalence. En outre, il est relativement difficile pour un algorithme de classement de candidats de décider de privilégier une référence de type NIL pour une mention, lorsque l'ensemble des candidats sélectionnés à l'étape précédente n'est pas vide. La proportion des références de type NIL assignées à bon escient sera donc très dépendante des résultats de l'étape de sélection des candidats. Ainsi, un score de rappel des références de type NIL élevé sera généralement annonciateur d'une proportion élevée de références de type NIL correctement assignées.

5. Mise en œuvre et résultats

Nous avons calculé les mesures présentées dans la section 5 pour différentes sources de données du web de données, envisagées comme ressources externes potentielles dans le cadre d'une application de résolution d'entités nommées de type «Personne». Ces calculs ont été réalisés à l'aide de requêtes SPARQL exécutées sur les points d'accès SPARQL du chapitre francophone de DBpedia (DBpFr), des bibliothèques nationales françaises (BNF) et espagnole (BNE) et des jeux de données de la fondation artistique Getty (Getty).

Une application de résolution d'entités nommées a été mise en œuvre à l'aide de l'outil REDEN (Brando *et al.*, 2015), dont le fonctionnement est conforme à la description des applications de résolution d'entités nommées non supervisées faite en section 3. Elle porte sur deux textes littéraires en français édités et fournis par le Labex OBVIL [12] que nous désignons ici par les noms de leurs auteurs, Apollinaire (A) (109 570 mots) et Thibaudet (T) (13 912 mots). Le premier traite de personnalités du monde de l'art vivant au début du 20e siècle, en particulier des artistes cubistes, des critiques et des marchands d'art. Le second a pour principal sujet des écrivains et des critiques littéraires du 19e siècle. Les annotations de référence utilisées pour nos tests ont été réalisées manuellement par des experts pour chacune des sources de données testées.

Le processus de résolution d'entités nommées est exécuté en prenant en entrée, pour le texte d'Apollinaire les sources de données Getty et DBpedia France, et pour celui de Thibaudet les sources de données DBpedia, BNE et BNF. Dans chaque cas, les mesures d'évaluation des résultats présentées en section 2 sont calculées. De plus, pour chaque texte, nous ajoutons des informations complémentaires afin de parfaire l'analyse des résultats : le nombre de mentions étiquetées (#TM), le nombre d'annotations manuelles (#Man), le nombre d'annotations de type NIL (#Nil), le nombre de mentions pour lesquelles la référence choisie est correcte (#CM), le nombre de mentions pour lesquelles la référence choisie est NIL et s'avère correcte (#CN).

Ces tests ont pour objectif d'évaluer si les éléments de la qualité des sources de données du web de données que nous avons identifiés et les mesures que nous avons proposées afin de les estimer quantitativement permettent ou non de décider quelle ressource externe est la plus pertinente pour une application de résolution d'entités nommées.

5.1. Évaluation de la qualité des sources de données du web de données

Les résultats du calcul des mesures d'évaluation de la qualité des quatre sources de données du web de données testées sont présentés dans le tableau 1.

Avec :

12. http://obvil.paris-sorbonne.fr

Tableau 1. Valeurs obtenues pour les mesures d'évaluation de la qualité de trois sources de données spécialisées et une source de données généraliste pour une application de résolution d'entités nommées de type «Personne» et «Personnes Françaises nées avant 1900»

Metric	Getty	BNE	BNF	DBpfr	BNF portée restreinte
$Port_{LD}$	199 521	480 892	**1 610 121**	207 429	67 488
PC_{LD}	256 (A)	1 095 (T)	1 880 (T)	1 121 (T) / 219 (A)	1 849 (T)
LR	2,9	**18,4**	1,4	1,6	1,8
$G1_{LD}$	**746**	-	63	24	47
$G2_{LD}$	**0,99**	-	0,04	0,92	0,18
PR_{LD}	0,3	0	0	**5,4**	0,9
EE_{LD}	0,2	**8,3**	2,5	0,9	2,5

– $Port_{LD}$: la portée d'un jeu de données,

– PC_{LD} : l'exhaustivité de population,

– LR : la richesse en termes de dénominations alternatives,

– $G1_{LD}$: la cardinalité de l'ensemble des concepts instanciés,

– $G2_{LD}$: la proportion des ressources qui instancient l'un des concepts plus spécialisés,

– PR_{LD} : présence de relations entre ressources du même type,

– EE_{LD} : l'exhaustivité des relations d'équivalence.

Le texte de Thibaudet est traité à l'aide de trois sources de données différentes, DBpedia, BNF et BNE. Les scores d'exhaustivité de population obtenus par les sources BNE et DBpedia sont relativement proches (respectivement 1095 et 1121). On note cependant que la source de données BNF a une exhaustivité de population considérablement meilleure (1 880) mais renferme un nombre de ressources extrêmement élevé correspondant aux critères de filtrage (1 610 121) par rapport aux deux autres sources de données (207 429 pour DBpedia et 480 892 pour BNE) et en moyenne moins de dénominations alternatives par ressource (1,4 pour BNF vs. 18,4 pour BNE et 16 pour DBpedia). On peut donc s'attendre, si on utilise la source de données BNF pour annoter les mentions du texte de Thibaudet à l'aide d'un processus de résolution d'entités nommées automatique non supervisé à avoir d'assez mauvais scores de résultats pour la phase de sélection des candidats, même si cette source de données contient plus de références valides que les deux autres sources envisagées. En effet, le dictionnaire généré à partir de la source de données BNF risque de comporter un grand nombre d'entrées non pertinentes par rapport à l'univers du discours du texte, mais homonymes des mentions du texte. Ceci risque donc d'engendrer un score de cardinalité moyenne des ensembles de candidats très élevé, un score de précision des candidats faible, un score de précision des références de type NIL élevé et un score de rappel des références de type NIL faible.

L'obtention de tels résultats lors de la phase de sélection des candidats laisse présager un effort de désambiguïsation important pour la phase de classement des candidats qui pourrait se solder par des résultats d'exactitude de désambiguïsation relativement mauvais. Avec moins de ressources vérifiant les critères de filtrage, mais des exhaustivités de population comparables et plus de dénominations alternatives, les sources de données BNE et DBpedia sont donc susceptibles d'obtenir des scores de cardinalité moyenne des ensembles de candidats plus faibles, de meilleurs scores de précision et de rappel des candidats et donc de meilleurs résultats d'exactitude de désambiguïsation lors de l'étape de classement. On note en particulier que seule la source de données DBpedia obtient des scores élevés pour les mesures de qualité liées au classement des candidats. Ainsi, pour s'assurer de meilleurs résultats pour cette étape du processus de résolution d'entités nommées pour le texte de Thibaudet, DBpedia devrait être utilisée en priorité. Lorsque nous restreignons la quantité de ressources issues de la source de données BNF et destinées à construire le dictionnaire de candidats en filtrant les données sur la base de critères de date de naissance et de langue, les mesures d'évaluation de la qualité de la source de données obtiennent de bien meilleurs scores, qu'il s'agisse des mesures liées à l'étape de sélection des candidats ou de celles liées à leur classement, tandis que la mesure d'exhaustivité de population conserve une valeur équivalente à celle estimée pour l'ensemble des données de type «Personne». On peut donc supposer qu'un processus de résolution d'entités nommées non supervisé appliqué sur le texte de Thibaudet obtiendrait de meilleurs résultats en utilisant ce jeu de données restreint plutôt que l'ensemble des ressources de type «Personne» disponibles sur le point d'accès SPARQL de la BNF.

Le texte d'Apollinaire est traité avec deux sources de données : Getty et DBpedia. Les deux ont des quantités de ressources candidates (199 521 pour Getty et 207 429 pour DBpedia) et des scores d'exhaustivité de population (256 pour Getty et 219 pour DBpedia) comparables. La base de connaissances Getty renferme des informations portant exactement sur le même thème que le texte à traiter et présente en moyenne 2,9 dénominations alternatives par ressource, là où DBpedia en propose seulement 1,6. On peut donc supposer que la mise en œuvre de ressources issues de cette source conduit à un meilleur score de rappel des candidats. En effet, plus on dispose de dénominations alternatives et plus il est facile de retrouver des candidats pertinents lorsque ceux-ci figurent bien dans la base de connaissances qui alimente le dictionnaire. De la même façon, le processus de résolution d'entités nommées sera plus susceptible de renvoyer des ensembles de candidats vides seulement dans les cas où il n'existe effectivement pas de référence valable dans la base de connaissances et pas en raison d'un manque de dénominations alternatives. Dans ces circonstances, le score de précision des références de type NIL devrait être meilleur. Les mesures de qualité associées à l'étape de désambiguïsation obtiennent de bons scores pour les deux bases de connaissances : les deux pourraient donc être mises à profit pour cette étape mais DBpedia se distingue néanmoins avec de meilleurs scores de présence de relations entre ressources du même type et devrait donc être privilégiée.

Les mesures d'évaluation de la qualité des sources de données du web de données calculées ici nous ont permis de proposer un avis sur les sources de données à

privilégier dans le cadre de deux applications de résolution d'entités nommées non supervisées. Nous allons maintenant nous attacher à confronter les conclusions tirées de l'analyse des résultats de ces mesures avec les scores d'évaluation des résultats réellement obtenus par un processus de résolution d'entités nommées appliqué à ces deux textes.

5.2. Comparaison des mesures d'évaluation a priori de la qualité des sources du web de données et des résultats effectivement obtenus par une application de résolution d'entités nommées non supervisée

Le tableau 2 présente les scores d'évaluation de résultats obtenus pour les textes de Thibaudet et d'Apollinaire par l'outil de résolution d'entités nommées REDEN (Brando *et al*., 2015) utilisant tour à tour les sources de données externes BNF (utilisée sur l'ensemble des ressources de type «Personne» et pour l'ensemble des ressources de type «Personne nées avant 1900 et de langue française» / BNF(pr)), BNE, Getty et DBpedia France.

Tableau 2. Résultats obtenus par un même processus de résolution d'entités nommées non supervisé appliqué sur deux textes et mettant en œuvre des données externes issues de sources spécialisées et généralistes

Crit.	A + Getty	A + DBpfr	T + BNE	T + BNF	T + DBpfr	T + BNF(pr)
#TM	297	297	3 404	3 404	3 404	3 404
#Man	257	266	1 850	1 880	1 849	1 916
#Nil	40	31	1 554	1 524	1 555	1 488
CardM	3,10	0,53	0,27	21,97	0,37	4,56
PrecCand	**0,96**	0,91	**0,59**	0,58	0,49	0,63
RapCand	**0,88**	0,52	0,26	**0,95**	0,32	**0,95**
PrecN	**0,55**	0,18	0,51	**0,91**	0,54	0,90
RapN	0,82	**0,84**	**0,85**	0,21	0,76	0,30
ExactD	0,82	**1**	**1**	0,65	**1**	0,88
ExactG	**0,82**	0,55	0,53	0,45	0,52	**0,62**
#CM	212	138	479	1 210	601	1 663
#CN	33	26	1 322	321	1 182	445

Avec :

– #TM : le nombre de mentions étiquetées,
– #Man : le nombre d'annotations manuelles,
– #Nil : le nombre d'annotations de type NIL,
– CardM : la cardinalité moyenne des ensembles de candidats,
– PrecCand : la précision des candidats,
– RapCand : le rappel des candidats,
– PrecN : la précision des références de type NIL,
– RapN : le rappel des références de type NIL,

– ExactD : l'exactitude de désambiguïsation,

– ExactG : l'exactitude globale,

– #CM : le nombre de mentions pour lesquelles la référence choisie est correcte,

– #CN : le nombre de mentions pour lesquelles la référence choisie est NIL et elle est correcte.

Les scores d'évaluation des résultats obtenus ici sont généralement cohérents avec les conclusions tirées des scores des mesures de qualité des jeux de données du web de données dans la sous-section précédente. Les tests réalisés sur le texte d'Apollinaire à l'aide des sources de données Getty et DBpedia, qui ont toutes deux obtenues de bons scores pour les mesures de qualité, obtiennent en moyenne de meilleurs scores pour chaque étape du processus de résolution d'entités nommées (une exactitude globale pour Apollinaire de 0,82). Ceci tend à confirmer que les mesures d'évaluation de la qualité des sources de données du web de données pour la résolution d'entités nommées dans des textes portant sur un domaine bien spécifique que nous avons proposées sont bien pertinentes et répondent à l'objectif visé d'aider au choix des sources de données externes les plus pertinentes pour ce type d'application. De plus, les bons scores obtenus par la base de connaissances Getty semblent confirmer l'importance du critère de portée du jeu de données utilisé en entrée du processus de résolution d'entités nommées. En effet, cette base a été produite pour décrire un domaine de connaissances bien particulier qui se trouve correspondre à l'univers du discours du texte traité, et donc elle renferme essentiellement des ressources pertinentes pour l'application à laquelle nous la destinons. Cette conclusion est aussi confirmée par les résultats bien meilleurs obtenus lors du traitement du même texte avec les données BNF dont la portée a été restreinte (une exactitude globale pour Thibaudet de 0,62) plutôt qu'avec l'ensemble des données BNF disponibles sur les ressources de type «Personne», car cela correspond essentiellement à une adaptation de la ressource au domaine du texte.

6. Conclusion et perspectives

Dans cet article, nous avons présenté un travail visant à proposer des mesures pertinentes pour l'évaluation *a priori* de la qualité de sources de données du web de données dans le cadre d'applications de résolution d'entités nommées non supervisées portant sur des textes spécialisés issus du domaine des humanités numériques. L'objectif de ces mesures est de fournir, autant que faire se peut, des indicateurs pertinents pour choisir, parmi les sources de données disponibles, celles pouvant conduire aux meilleurs résultats de résolution d'entités nommées possibles pour un même texte et un même outil de traitement. Nous avons présenté le déroulement global d'un processus classique de résolution d'entités nommées non supervisé et en avons déduit les principaux éléments de qualité à privilégier pour la sélection des données externes à mettre en œuvre dans un tel processus. Nous avons ensuite sélectionné et adapté des mesures d'évaluation de la qualité de sources de données du web de données de la littérature afin de permettre une évaluation quantitative de ces éléments de qualité. Puis, nous avons calculés les valeurs de ces mesures pour quatre sources de données poten-

tiellement utilisables pour annoter deux textes littéraires en français. Enfin nous avons confronté les conclusions tirées de ces valeurs avec les scores d'évaluation des résultats effectivement obtenus par un processus de résolution d'entités nommées appliqué sur ces textes avec les différents jeux de données envisagés. Les résultats obtenus semblent prouver la pertinence des mesures proposées.

Cependant leur mise en œuvre, faite à l'aide de requêtes SPARQL, demeure manuelle en raison de la difficulté d'automatisation de la création des requêtes utiles. En effet, d'une part, le choix des concepts et prédicats pertinents pour filtrer les données en entrée du dictionnaire, et d'autre part la prise en compte des choix de structurations de données et de vocabulaires différents d'une source de données à l'autre restent des tâches complexes à automatiser. L'approche proposée par Nikolov *et al*. (2011) pour automatiser l'identification des sources de données du web les plus pertinentes auxquelles lier un jeu de données à publier sur le web de données, et au sein de ces sources, les classes à utiliser pour filtrer les ressources candidates au liage semble néanmoins adaptable à notre problématique d'identification de sources de données pertinentes et d'identification automatique de critères de filtrage des ressources à inclure dans le dictionnaire des candidats.

Cependant, l'approche de définition et d'évaluation d'éléments de qualité des sources de données du web de données pertinents pour une application donnée que nous avons suivie ici semble généralisable pour d'autres types d'applications. Ainsi, on pourrait envisager de publier avec les métadonnées des jeux de données publiés sur le web, des valeurs de mesures d'évaluation de la qualité de ces jeux de données pour diverses applications. L'intérêt de cette démarche d'auto-évaluation de la qualité d'une jeu de données pour certains types d'applications est double pour les producteurs de données. D'une part, elle permet de mieux appréhender l'adéquation réelle des données publiées aux besoins des différentes applications susceptibles de les utiliser et ainsi d'identifier et de pallier certains manques éventuels de ces données. D'autre part, elle favorise leur réutilisation par les applications pour lesquelles elles s'avèrent les plus adaptées en signalant directement aux développeurs leur degré d'adéquation aux besoins.

Remerciements

Les auteurs remercient le Labex OBVIL[13] *de l'Université Paris-Sorbonne pour leur avoir fourni les textes annotés utilisés pour les expériences.*

Bibliographie

Besançon R., Daher H., Ferret O., Le Borgne H. (2016). Utilisation des relations d'une base de connaissances pour la désambiguïsation d'entités nommées. In *23ème conférence sur le traitement automatique des langues naturelles (jep-taln-recital 2016)*, p. 290–303.

Brando C., Frontini F., Ganascia J. (2016). REDEN: named entity linking in digital literary editions using linked data sets. *CSIMQ*, vol. 7, p. 60–80. Consulté sur http://dx.doi.org/10.7250/csimq.2016-7.04

Brando C., Frontini F., Ganascia J.-G. (2015). Disambiguation of named entities in cultural heritage texts using linked data sets. In *New trends in databases and information systems*, vol. 539, p. 505-514. Springer.

Cheatham M., Hitzler P. (2013). String similarity metrics for ontology alignment. In *International semantic web conference*, p. 294–309.

Erp M. van, Mendes P., Paulheim H., Ilievski F., Plu J., Rizzo G. *et al*. (2016, 05). Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *LREC 2016, 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Portoroz, Slovenia.* Portoroz, Slovénie. Consulté sur http://www.eurecom.fr/publication/4859

Ferraram A., Nikolov A., Scharffe F. (2013). Data linking for the semantic web. *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications*, vol. 169.

Gruetze T., Kasneci G., Zuo Z., Naumann F. (2016). Coheel: Coherent and efficient named entity linking through random walks. *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 37, n° 0. Consulté sur http://www.websemanticsjournal.org/index.php/ps/article/view/463

Hachey B., Radford W., Curran J. R. (2011). Graph-based named entity linking with wikipedia. In A. Bouguettaya, M. Hauswirth, L. Liu (Eds.), *Wise*, vol. 6997, p. 213-226. Springer. Consulté sur http://dblp.uni-trier.de/db/conf/wise/wise2011.html#HacheyRC11

Han X., Sun L., Zhao J. (2011). Collective entity linking in web text: A graph-based method. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval*, p. 765–774. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/2009916.2010019

Jentzsch A., Mühleisen H., Naumann F. (2015). Uniqueness, density, and keyness: Exploring class hierarchies. In *Proceedings of the 6th international workshop on consuming linked data (cold 2015).* Consulté sur http://ceur-ws.org/Vol-1426/paper-03.pdf

Mihalcea R., Csomai A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management*, p. 233–242. New York, NY, USA, ACM. Consulté sur http://doi.acm.org/10.1145/1321440.1321475

Moro A., Raganato A., Navigli R. (2014). Entity linking meets word sense disambiguation: a unified approach. *TACL*, vol. 2, p. 231-244. Consulté sur http://dblp.uni-trier.de/db/journals/tacl/tacl2.html\#0001RN14

Nikolov A., d'Aquin M., Motta E. (2011). What should i link to? identifying relevant sources and classes for data linking. In *Joint international semantic technology conference*, p. 284–299.

Paulheim H., Bizer C. (2014, avril). Improving the quality of linked data using statistical distributions. *Int. J. Semant. Web Inf. Syst.*, vol. 10, n° 2, p. 63–86. Consulté sur http://dx.doi.org/10.4018/ijswis.2014040104

Pustejovsky J. (1991). The generative lexicon. *Computational linguistics*, vol. 17, n° 4, p. 409–441.

Ruckhaus E., Vidal M., Castillo S., Burguillos O., Baldizan O. (2014). Analyzing linked data quality with liquate. In *The semantic web: ESWC 2014 satellite events - ESWC 2014 satellite events, anissaras, crete, greece, may 25-29, 2014, revised selected papers*, p. 488–493. Consulté sur http://dx.doi.org/10.1007/978-3-319-11955-7_72

Schmidt M., Lausen G. (2013). Pleasantly consuming linked data with rdf data descriptions. *CoRR*, vol. abs/1307.3419. Consulté sur http://dblp.uni-trier.de/db/journals/corr/corr1307.html\#SchmidtL13

Shvaiko P., Euzenat J. (2005). A survey of schema-based matching approaches. In *Journal on data semantics iv*, p. 146–171. Springer.

Sinha R. S., Mihalcea R. (2007). Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *Icsc*, vol. 7, p. 363–369.

Usbeck R., Ngomo A.-C. N., Röder M., Gerber D., Coelho S., Auer S. *et al.* (2014). AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data. In P. Mika *et al.* (Eds.), *The semantic web – iswc 2014*, vol. 8796, p. 457-471. Springer International Publishing. Consulté sur http://dx.doi.org/10.1007/978-3-319-11964-9_29

Zaveri A., Rula A., Maurino A., Pietrobon R., Lehmann J., Auer S. (2015). Quality assessment for linked data: A survey. *Semantic Web Journal*. Consulté sur http://www.semantic-web-journal.net/content/quality-assessment-linked-data-survey