Atelier Qualité des Données du Web (QLOD'16)

Organisateurs : Samira Si-said Cherfi, Fayçal Hamdi (CEDRIC - Cnam Paris)

PRÉFACE

Dans le contexte actuel de forte compétitivité, les organisations de tous horizons sont face à un nouveau défi qui est celui de la valorisation des données. Ces données qui, grâce au développement des technologies du web, deviennent un atout stratégique. La disponibilité des données dans un format numérique et structuré a accéléré le développement des techniques d'exploration et de raisonnement sur les données dans le but d'encourager l'innovation et d'extraire de la valeur ajoutée. Ces approches ont créé un besoin pour des données réelles, fiables et de haute qualité. En conséquence, la qualité des données est devenue un enjeu important et un défi pour les années à venir.
La recherche dans ce domaine comprend des aspects théoriques, liés à la formalisation, la définition de la qualité et le développement de langages et de modèles supportant les concepts sous-jacents. Elle couvre des recherches pratiques et expérimentales par exemple le développement de méthodes d'évaluation, la proposition d'approches de validation et de benchmarks, l'expérimentation sur des jeux réels, etc.

QLOD vise à offrir un espace pour des échanges fructueux et enrichissant impliquant des chercheurs, mais aussi des professionnels du monde de l'entreprise avec une diversité de point de vue et de profil concernant la qualité des données surtout lorsqu'elles sont hétérogènes, non/peu structurées, massives et de qualités diverses.

Cette première édition présente trois travaux sur le lien entre les données et les connaissances qu'elles véhiculent, sur la qualité des données géolocalisées et sur la qualité des données liées. Ces thèmes sont complétés par une présentation de Fabian M. Suchanek autour de la construction des ontologies et de l'intégration des sources de données avec les problèmes et les opportunités de recherche sous-jacents.

Samira Si-said Cherfi
CEDRIC - Cnam Paris

Fayçal Hamdi
CEDRIC - Cnam Paris

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Nathalie Abadie
Jacky Akoka
Saïd Assar
Isabelle Comyn-Wattiau
Jérôme David
Virginie Goasdoue-Thion
Zoubida Kedad
Benjamin Nguyen
Verónika Peralta
Jolita Ralyté
Chantal Reynaud
Fatiha Saïs
Grégory Smits

TABLE DES MATIÈRES

Mise en correspondance de données textuelles hétérogènes à partir d'informations sémantiques

Nourelhouda YAHI*, Hacene BELHADEF*, Mathieu ROCHE**

* Constantine 2 Abdelhamid Mehri University, Nouvelle Ville Ali Mendjeli, Constantine, Algeria
** UMR TETIS (Cirad, Irstea, AgroParisTech) & LIRMM (CNRS, Univ. Montpellier), France

Résumé : Dans cet article, nous présentons une approche pour mesurer la similarité sémantique entre des textes hétérogènes et de qualité différente provenant de différentes sources Web. Notre approche commence par extraire le contenu des textes par deux méthodes : (i) utilisation d'un système d'extraction que nous avons implanté et qui identifie tous les mots contenus dans un texte donné, (ii) utilisation d'un thésaurus multilingue (AGROVOC). Ensuite, nous combinons les résultats des deux approches afin de mesurer la similarité entre les représentations textuelles des documents. Afin d'évaluer les résultats, nous nous appuyons sur deux ensembles de données hétérogènes issus du Web (tweets et articles scientifiques).

1. Introduction

Pour traiter les masses de données issues du Web disponibles, la problématique de recherche du *Big Data* est classiquement mise en avant avec les 3V qui la caractérisent : volume, variété et vélocité. Même si une distinction est établie entre la *véracité* (qualité) et la *variété* (hétérogénéité) des données, l'imbrication de ces deux concepts doit être prise en compte. En effet, pour avoir une connaissance exhaustive d'un sujet donné, il est nécessaire de traiter et de mettre en relation les données hétérogènes et de qualité différente. Ceci améliore indéniablement ce que nous pouvons appeler *la qualité des connaissances* traitée en prenant en compte différents points de vue. En effet, dans le domaine des SHS, une problématique tout à fait ouverte consiste à analyser des situations en considérant les multiples points de vue à travers les dires d'acteurs et d'experts. Par exemple, lorsque l'on parle de changement climatique (jeux de données traité dans cet article) plusieurs positions peuvent être mises en relief sur des supports différents (articles de presse, tweets, articles scientifiques, etc.). La qualité des connaissances est donc fortement liée à la diversité des points de vue abordés sur un sujet donné. Le lien entre *qualité des données* et *hétérogénéité* de ces dernières est donc important à considérer. Dans ce contexte, nous nous intéressons à la manière de mettre en correspondance des données textuelles hétérogènes qui sont, par nature, de qualité diverse (formats, contenus, styles linguistiques, etc.).

Plusieurs projets de recherche s'intéressent à la similarité sémantique entre des extraits de textes, mais la plupart d'entre eux s'appuient sur des textes ayant un même « niveau » linguistique et stylistique (Maguitman *et al.*, 2005) (Elsayed et Oard, 2008). Le développement d'une approche efficace qui permet de proposer une similarité sémantique entre les textes hétérogènes représente alors une problématique eminament difficile. Il existe un certain nombre de travaux de la littérature liés à l'estimation de similarité sémantique, dont beaucoup sont fondés sur l'utilisation de thésaurus. Par exemple, (Buscaldi *et al.*, 2012) proposent un processus de comparaison de n-grammes sur la base d'une mesure de similarité conceptuelle utilisant WordNet[1]. Ils ont aussi appliqué une démarche similaire pour calculer la similarité sémantique de fragments textuels.

Les textes peuvent être écrits selon des styles très différents, par exemple, les tweets sont beaucoup plus difficiles à analyser linguistiquement. *A contrario*, les articles scientifiques ont une écriture plus standardisée permettant l'extraction d'information de manière plus aisée. Mais ce type de textes possède un vocabulaire de spécialité souvent plus complexe (Batista-Navarro *et al.* 2015). Dans cet article, nous proposons l'approche MIGHT (A Text Mining Process for Mapping Heterogeneous Documents) pour mesurer la similarité sémantique entre les textes hétérogènes et de qualité différente. Notre approche utilise un système extraction que nous avons implanté et qui s'appuie sur le thésaurus multilingue AGROVOC. Nous avons alors combiné les informations extraites pour calculer la similarité entre les représentations textuelles. L'article présente notre approche (section 2) qui est évaluée sur un jeu de données réel sur la thématique du changement climatique (section 3).

2. Approche MIGHT

Dans cette section, nous décrivons les détails de l'approche proposée consistant à mesurer la similarité sémantique entre deux textes de qualité différente.

Avant de mesurer la similarité entre les textes, une pondération des descripteurs linguistiques est classiquement mise en place. La pondération des termes permet d'identifier leur importance dans le texte. En général, l'idée de base est d'attribuer des poids aux termes en utilisant des informations statistiques telle que la fréquence dans un texte ou relativement à un corpus dans son ensemble (TF-IDF, Okapi, etc.).

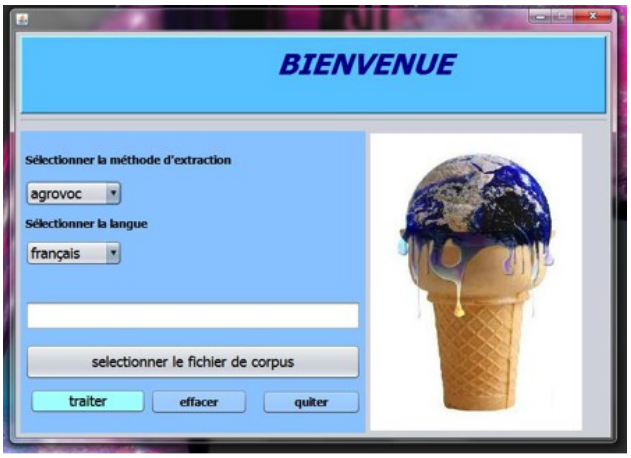
Dans notre approche, étant donné que nous traitons des textes très différents (des textes longs mais également très courts), nous avons adopté une pondération différente : nous cherchons à donner un poids plus élevé à des termes véhiculant une certaine sémantique

[1] https://wordnet.princeton.edu/

(illustré par leur appartenance à la ressource AGROVOC). Les thésaurus sont largement utilisés pour l'estimation de similarité comme WordNet qui modélise la connaissance lexicale en anglais. Dans notre approche, nous utilisons AGROVOC[2], thésaurus multilingue du domaine agronomique, qui couvre tous les domaines d'intérêt de la FAO, Organisation des Nations Unies pour l'alimentation et l'agriculture. Il est publié par la FAO et édité par une communauté d'experts. AGROVOC se compose de plus de 32.000 concepts disponibles en 23 langues. À ce jour, AGROVOC est utilisé par les chercheurs, les bibliothécaires et les gestionnaires de l'information pour l'indexation, l'extraction et l'organisation des données dans les systèmes d'information agricoles (Roche *et al.* 2015).

Les pondérations des descripteurs linguistiques ont été réalisées de la façon suivante : tous les mots identifiés dans un texte par rapport à une base représentant l'ensemble des descripteurs des corpus sont pondérés à 1, les termes qui sont extraits avec l'extracteur d'AGROVOC sont pondérés à 2 et les descripteurs linguistiques identifiés sur la base de ces deux méthodes sont pondérés à 3.

Afin d'évaluer l'approche proposée, nous avons développé un logiciel dédié (cf. figure ci-dessous).



La première étape du processus supprime tous les signes de ponctuation issus des données textuelles, puis élimine tous les « stop-words » afin d'extraire les mots, *a priori*, porteurs d'information sémantique. La deuxième étape est fondée sur l'extraction de termes

[2] http://aims.fao.org/fr/agrovoc

(mots et syntagmes) avec AgroTagger[3] ; pour cette tâche, nous nous sommes appuyés sur une classe spécifique (Maui). Le résultat de ces deux tâches est stocké dans une base de données et des vecteurs relatifs à chaque document sont construits. Enfin, une mesure de similarité (cosinus) calcule la proximité entre deux textes donnés (vecteurs pondérés).

3. Expérimentations

Dans cette section, nous montrons comment appliquer notre approche sur des ensembles de données hétérogènes (tweets et articles scientifiques) en langue française relativement à la thématique « changement climatique ».

Corpus et protocole expérimental

Dans ces expérimentations, nous avons d'abord recueilli des tweets en suivant sur les comptes Twitter des hashtags spécifiques : *#réchauffementClimatique*, *#changementClimatique*. Ainsi, nous avons constitué un corpus de tweets français issus d'associations, d'organisations, de célébrités et de citoyens abordant cette thématique. Puis trois autres corpus ont été utilisés. Le premier, Politweets (Longhi *et al.*, 2014), rassemble des tweets de 7 personnalités issus de 6 différents groupes politiques français (34273 messages). Le deuxième corpus est une collection de résumés en français d'articles, de livres, de chapitres de livres, de thèses, etc., à partir Agritrop[4], archive ouverte du Cirad (Centre de coopération internationale en recherche agronomique pour le développement). Ces résumés scientifiques traitent du sujet du changement climatique. Le dernier corpus est une collection de résumés en français d'articles issus du laboratoire TETIS (Territoires, Environnement, Télédétection et Information Spatiale).

Ces 4 collections de données textuelles disponibles sur le Web sont notées de la manière suivante :

- CT : Tweets traitant du changement climatique.
- NCT : Tweets non liés au changement climatique (*Politweets*).
- CA : Articles scientifiques traitant du changement climatique (*Cirad*).
- NCA : Articles scientifiques non liés au changement climatique (*TETIS*).

[3] http://aims.fao.org/vest-registry/tools/agrotagger-1
[4] https://agritrop.cirad.fr/

Résultats

Les expérimentations sont réalisées selon 10 itérations, pour chacune d'elle, nous avons sélectionné aléatoirement N éléments (N = 10) – le nombre d'exécutions total est de 3000. Ensuite, nous appliquons la similarité entre les éléments (les documents).

Le tableau ci-dessous montre les résultats obtenus (moyenne des similarités pour chaque itération) afin de comparer (i) CT et CA, (ii) CT et NCT, (iii) CT et NCA. Les degrés de similarité les plus élevés pour cinq itérations est la similitude entre CT et CA (couvrant des sujets proches) mettent en avant des premiers résultats encourageants restitués par notre approche MIGHT. Dans la suite des travaux, il sera nécessaire d'analyser d'un point de vue qualitatif les résultats obtenus, en particulier les faux positifs et faux négatifs obtenus. Ceci permettra de discuter dans quelle mesure les mots originaux issus des textes ou les termes d'Agrovoc ont permis d'améliorer ou non les différentes mises en correspondance des documents.

Iteration	CT/NCT	CT/CA	CT/NCA
1	0.0073	0.0025	0.0078
2	0	0.0063	0
3	0	0.0128	0.0093
4	0	0	0
5	0	0	0
6	0	0.0182	0.032
7	0	0.0069	0
8	0	0.0182	0.032
9	0	0.0069	0
10	0	0.0106	0

4. Conclusion

Dans cet article, nous avons abordé la question de l'évaluation de la similarité sémantique entre des documents de nature différente mais qui peuvent porter sur des sujets proches. Notre approche est fondée sur l'extraction de descripteurs linguistiques issus d'un texte (mots) et des termes (mots et syntagmes) propres à un thésaurus en appliquant une pondération « sémantique » spécifique. Notre méthode a tendance à rapprocher des textes ayant des thématiques proches ce qui permet de mettre en relation des données de qualité différente. De nombreuses perspectives peuvent être proposées comme (i) l'élimination du vocabulaire spécifique aux tweets (phrases ou expressions spécifiques), (ii) l'expansion de contextes (par exemple, en considérant l'ensemble de tweets écrits par le même auteur dans une même fenêtre temporelle).

Remerciements

Ce travail est financé par la Région Languedoc-Roussillon et par les Fonds Européens de Développement Régional (**projet SONGES** : http://textmining.biz/Projects/Songes).

Références

R. T. Batista-Navarro, R. Rak, S. Ananiadou (2015). Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminformatics 7(S-1): S6*

D. Buscaldi, R. Tournier, N. Aussenac-Gilles, J. Mothe (2012). Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In F*irst Joint Conference on Lexical and Computational Semantics (SEM)*, pages 552-556, Montreal, Canada.

T. Elsayed, J. Lin, D. W. Oard (2008). Pairwise document similarity in large collections with mapreduce. In P*roceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*: Short Papers, Columbus, Ohio, 2008, pp.2652-68.

J. Longhi, C. Marinica, B. A. Alkhouli (2014). Polititweets : corpus de tweets provenant de comptes politiques influent. In *Chanier T. Banque de corpus CoMeRe*. Ortolang.fr.

A.G. Maguitman, F. Menczer, H. Roinestad, A. Vespignani (2005). Algorithmic detection of semantic similarity. In *Proceedings of 22nd International Conference on World Wide Web*, Chiba, Japan, 2005, pp. 107-116.

M . Roche, S. Fortuno, J.A. Lossio-Ventura, A. Akli, S. Belkebir, T. Lounis, S. Toure (2015). Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cahiers Agricultures*. Volume 24, numéro 5, p.313-320

Que représentent les références spatiales des données du Web ? un vocabulaire pour la représentation de la sémantique des XY

Abdelfettah FELIACHI*, Nathalie ABADIE*
Fayçal HAMDI**,

*IGN-COGIT, Saint-Mandé, France.
abdelfettah.feliachi@ign.fr, nathalie-f.abadie@ign.fr
**CNAM-CEDRIC, Paris, France.
faycal.hamdi@cnam.fr

Résumé. La référence spatiale qui décrit la localisation des ressources Linked Data peut représenter un critère très utile pour l'interconnexion de ces ressources. Cependant, dans le contexte du Web de données cette référence spatiale peut être très hétérogène entre les différentes sources de données. Nous aborderons dans cet article la question de l'identification des causes de ces hétérogénéités et nous proposerons un modèle de représentation de ces causes afin de pouvoir les gérer de manière plus efficace dans le processus d'interconnexion.

1 Contexte et objectif

De plus en plus de ressource publiées dans le Web des données possèdent dans leur description une référence spatiale qui représente leur localisation dans le monde réel d'une manière directe ou indirecte. Dans le processus d'interconnexion des ressources des différentes sources du Web des données, ces références spatiales représentent l'un des critères privilégiés. En effet, deux ressources proches spatialement sont très susceptibles de représenter la même entité du monde réel. Ce principe est appliqué aussi bien dans le domaine de l'appariement de données géographiques que dans celui de l'interconnexion de Linked Data géoréférencées. Les références spatiales sont généralement comparées par le calcul d'une distance géographique dans le cas d'un référencement direct, ou par des distances de chaînes de caractères dans le cas d'un référencement indirect. Cependant la représentation des références spatiales peut être très hétérogène d'une source à une autre et parfois même parmi les ressources issues d'une même source. Ceci peut être compromettant pour l'utilisation de ce critère. Cette hétérogénéité des références spatiales peut être dû au fait que la référence spatiale dans le Web des données est souvent utilisée pour localiser la ressource qu'elle décrit d'une manière simple : par des adresses postales, des coordonnées de longitude et de latitude ou des géométries de type "point". Le cas des bases de données géographiques classiques est différent : leur référence spatiale est représentée de manière détaillée (point, ligne ou polygone par exemple). La référence spatiale n'a pas le même niveau d'importance, et donc pas la même qualité, entre

le Web des données et les bases de données géographiques classiques. Les causes d'hétérogénéité des références spatiales sont généralement peu prises en compte, et seulement d'une manière implicite, dans les processus d'interconnexion. Nous pensons qu'une identification de ces causes d'hétérogénéité dans le contexte du Web des données et une explicitation de celles-ci peuvent améliorer leur prise en compte dans les processus d'interconnexion. Donc, notre objectif est de proposer un modèle pour expliciter la sémantique des références spatiales afin de fournir des informations exploitables dans le processus d'interconnexion.

2 Un vocabulaire pour décrire les causes d'hétérogénéité entre références spatiales directes

En s'appuyant sur les causes d'hétérogénéité des données géographiques décrites dans la littérature, notamment les sources de conflits d'intégration de données géographiques décrites par Devogele (1997) et les sources d'imperfections de données géographique décrites par Girres (2012), on distingue plusieurs types et causes d'hétérogénéité entre les références spatiales des différentes sources. Dans le contexte du Web de données, on peut identifier entre autres : l'utilisation de différents types de références spatiales (directe ou indirecte), l'utilisation de différents vocabulaires pour la représentation des références spatiales, la différence dans les sources d'information, le processus de saisie ou d'extraction, la modélisation géométrique et le niveau de précision géométrique, etc. Dans le cadre de l'interconnexion des données géoréférencées, nous nous intéressons plus particulièrement aux trois éléments suivants, que nous jugeons plus susceptibles d'influencer le processus d'interconnexion :

- La différence de précision géométrique des ressources.
- La différence dans la modélisation géométrique des références spatiales des ressources.
- L'aspect vague des entités géographiques que les ressources visent à représenter.

Nous proposons donc un modèle qui permet de représenter ces informations dans un jeu de données au niveau des références spatiales des ressources, c'est-à-dire de leur géométrie. Nous choisissons de représenter ces informations au niveau le plus fin, à savoir la géométrie de chaque ressource. En effet ces informations peuvent changer d'une ressource à une autre, même au sein d'un seul jeu de données. De plus, une ressource peut avoir dans sa description plusieurs références spatiales nécessitant chacune des métadonnées différentes.

En l'absence de standard encore bien installé pour la représentation des géométries sur le Web des données, nous proposons d'étendre l'ontologie des géométries [1] qui présente l'avantage d'être compatible avec GeoSPARQL et de permettre la représentation de géométries structurées Hamdi et al. (2014). Nous présentons désormais les choix de modélisation effectués.

2.1 La précision géométrique

La précision planimétrique est le critère de qualité de données géographiques qui représente les écarts entre les positions des objets géographiques et les positions des entités réelles qu'elles représentent. Nous nous appuyons sur la description des éléments de qualité des données géographiques tels que définis dans les normes ISO19157 (2013) et ISO19115 (2003),

1. `http://data.ign.fr/def/geometrie`

afin de représenter l'élément de précision planimétrique. Cet élément est décrit principalement par le type de sa méthode d'évaluation et le résultat quantitatif de cette évaluation (Fig.1).

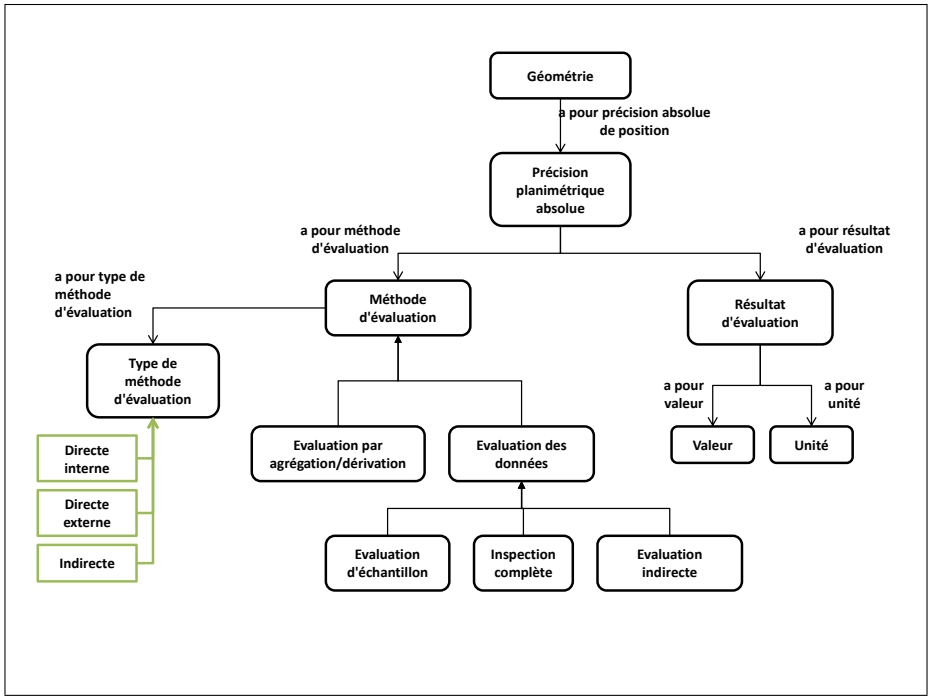


FIG. 1 – *Extrait du modèle proposé de la représentation de l'élément de qualité qui décrit la précision planimétrique d'une référence spatiale.*

Un exemple de l'utilisation possible de cette partie du modèle afin de décrire la précision planimétrique d'une référence spatiale est illustré dans la figure suivante (Listing 1).Dans cet exemple la précision planimétrique est de 100 mètres. Elle a été estimée par une vérification directe sur échantillon de données par rapport à la réalité du terrain.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>
@prefix geom:<http://data.ign.fr/def/geometrie#>.
@prefix xysemantics:<http://data.ign.fr/def/xysemantics#>.

<http://data.ign.fr/id/xx/xx/xx> geom:geometry <http://data.ign.fr/id/xx/xx/
     Multipolygon_XX>.

<http://data.ign.fr/id/xx/xx/Multipolygon_xx> a geom:MultiPolygon;
xysemantics:absolutePositionalAccuaracy [
    a xysemantics:AbsolutePositionalAccuaracy;
    xysemantics:evaluationResult [
        a xysemantics:Result;
        xysemantics:value "100"^^xsd:float;
        <http://qudt.org/schema/qudt#unit> <http://qudt.org/vocab/unit#Meter>];
    xysemantics:evaluationMethod [
        a xysemantics:SampleBasedInspection;
```

```
    xysemantics:evaluationMethodTypeCode <http://data.ign.fr/id/codes/geomtrie/
        methodeevaluation/DirectExternal>]].
```

Listing 1 – *Exemple de triplets RDF qui représentent la précision planimétrique*

2.2 La modélisation géométrique

Elle représente le choix de représentation géométrique par rapport à l'entité du monde réel référencée par l'objet géographique. Gesbert (2005) et Abadie (2012) se sont intéressés à la description formelle de la modélisation géométrique dans le cadre de leurs travaux de modélisation des règles de saisie des données géographiques, afin d'améliorer leur intégration. Dans ce cadre, nous notons principalement la proposition de Abadie (2012) basée sur les travaux de Sahade (2010) et Smith et Mark (1998), qui définit un ensemble d'éléments représentatifs de la forme des entités géographiques du monde réel, tels que perçus dans un contexte cartographique. Ceci est concrétisé au niveau de notre modèle par un « élément caractéristique de la forme » décrit par son type, qui fait référence à une taxonomie des types d'éléments caractéristiques de la forme (Fig.2). Cette taxonomie peut éventuellement être étendue. Ces éléments de forme sont des éléments parasites qui ne peuvent exister sans un objet hôte, qui est l'entité du monde réel qui les porte (Fig.2).

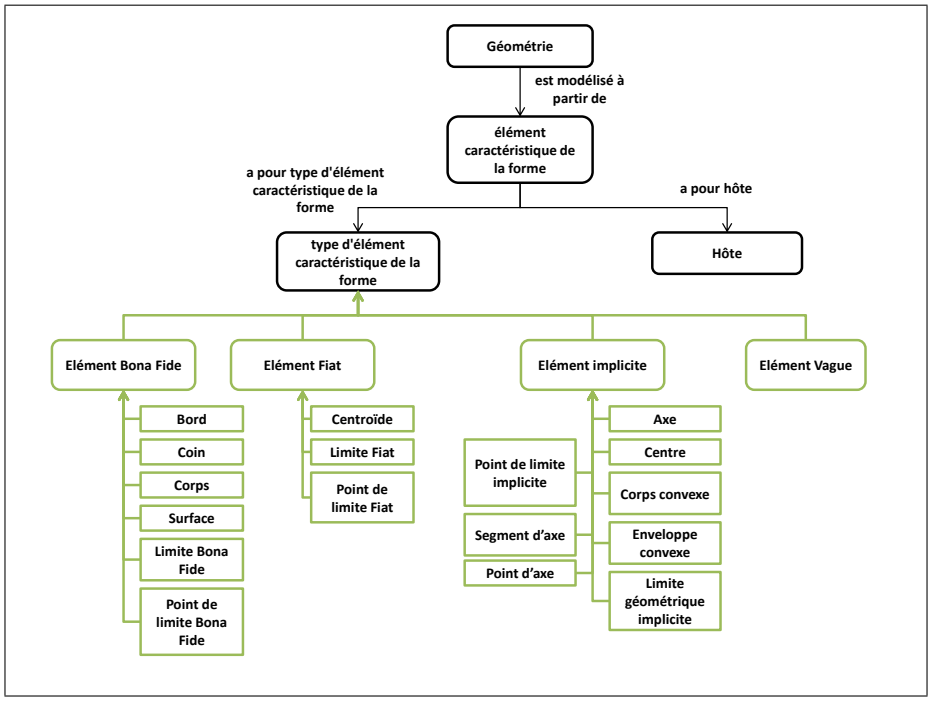


FIG. 2 – *Extrait du modèle proposé qui décrit l'élément de représentation géométrique d'une ressource spatiale.*

Un exemple possible de l'utilisation de cette partie de modèle afin de décrire la modélisation géométrique d'une référence spatiale est illustré dans la figure suivante (Listing 2). Dans cet exemple la géométrie représente une limite de type Bona Fide saisie à partir des bâtiment.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix geom:<http://data.ign.fr/def/geometrie#>.
@prefix xysemantics:<http://data.ign.fr/def/xysemantics#>.

<http://data.ign.fr/id/xx/xx/xx> geom:geometry <http://data.ign.fr/id/xx/xx/
    Multipolygon_XX>.

<http://data.ign.fr/id/xx/xx/Multipolygon_xx> a geom:MultiPolygon;
xysemantics:sModeledFrom [
    a xysemantics:ShapeCharacteristicElement;
    xysemantics:shapeCharacteristicElementType <http://data.ign.fr/id/codes/geometrie/
        elementcaracteristique/BonaFideBoundary>;
    <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#host> <http://data.ign.fr/def/topo#
        Bati> ].
```

Listing 2 – *Exemple de triplets RDF qui représentent la modélisation géométrique*

2.3 L'aspect vague des entités géographiques

Certaines entités géographiques sont difficiles à définir ou délimiter par essence, comme c'est le cas d'une vallée par exemple. Dans ce cas, l'élément caractéristique de la forme géométrique doit être de type « élément vague » (Fig.2).

3 Conclusion

Dans cet article nous nous sommes intéressés aux causes des hétérogénéités entre les ressources géoréférencées qui peuvent affecter la fiabilité du critère spatial dans le processus d'interconnexion. Nous identifions un ensemble d'éléments permettant une représentation plus riche de la précision planimétrique et de la modélisation géométrique de chaque référence spatiale. Nous avons organisé ces éléments dans un modèle sous forme d'ontologie [2] OWL. La suite de notre travail consistera à mettre au point une approche exploitant ce modèle afin d'améliorer la prise en compte du critère spatial dans un processus d'interconnexion.

Références

Abadie, N. (2012). *Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques : Les spécifications au cœur du processus d'intégration*. Thèse de doctorat, Université Paris-Est.

Devogele, T. (1997). *Processus d'intégration et d'appariement des bases de données géographiques. Application à une base de données routière multi-échelles*. Thèse de doctorat, Université Paris-Est.

2. http ://data.ign.fr/def/xysemantics

Gesbert, N. (2005). *Formalisation des spécifications de bases de données géographiques en vue de leur intégration*. Thèse de doctorat, Université Paris-Est.

Girres, J.-F. (2012). *Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Application aux mesures de longueur et de surface*. Thèse de doctorat, Université Paris-Est.

Hamdi, F., N. Abadie, B. Bucher, et A. Feliachi (2014). Geomrdf: A geodata converter with a fine-grained structured representation of geometry in the web. *The 1st International Workshop on Geospatial Linked Data - (GeoLD 2014) - SEMANTiCS 2014*, pp.12.

ISO19115 (2003). Geographic information – metadata. Standard, International Organization for Standardization (TC 211).

ISO19157 (2013). Geographic information – data quality. Standard, International Organization for Standardization (TC 211).

Sahade, S. (2010). Computer-tractable translation of geospatial data. *International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission) 5*.

Smith, B. et D. Mark (1998). Ontology and geographic kinds. *8th International Symposium on Spatial Data Handling (SDH'98)*, 308–320.

Summary

The spatial reference that describes the location of Linked Data resources can be a very useful criterion for the interlinking of these resources. However, in the context of Web of data this spatial reference can be very heterogeneous between different data sources. We discuss in this article the question of identifying the causes of these heterogeneities and we propose a model of representation of these causes in order to better manage them in the interlinking process.

Linked Data Quality for Domain-Specific Named-Entity Linking

Carmen Brando*, Nathalie Abadie**, Francesca Frontini***

*IGN, ValiLab-DUIG, Marne-la-Vallée, France
Carmen.Brando@ign.fr

**Université Paris-Est, IGN/SRIG, COGIT, Saint-Mandé, France
Nathalie-F.Abadie@ign.fr

***Istituto di Linguistica Computazionale CNR, Pisa, Italy
Francesca.Frontini@ilc.cnr.it

Résumé. We present outgoing research whose goal is to assess quality of Linked Data for its usage in domain-specific Named-entity Linking (NEL). NEL is the task of assigning appropriate referents, typically an Uniform Resource Identifier (URI), to mentions of entities (e.g. persons or places) identified in textual documents. Nowadays, many of these approaches strongly rely on Linked Data as knowledge base. However, the scope of the chosen data sets can have an important influence on the performances of NEL as texts often concern specific domains of knowledge. In this paper, we describe LD quality aspects which should be considered for improving NEL in domain-specific contexts, then propose quality metrics and compute them for both French DBpedia and the French National Library (BnF) data sets thereby to discuss the opportunity of using these data sets for the linking of authors in old French Literary digital editions. Our ultimate goal is to improve a Natural Language Processing (NLP) pipeline for the automatic annotation of these texts.

1 Introduction

Data quality is often referred to as fitness for use, in other words, the capability of a data set to fit the requirements for a given use case or for performing a particular task. The increasing volume and availability of Linked Data (LD) brings new opportunities to build LD-based tools. In this context, LD sets such as DBpedia, Pleiades [1] or the French National Library (BnF) [2] serve frequently as external resources to Natural Language Processing (NLP) tasks such as Named-Entity Linking (NEL). In general terms, NEL aims to assign the appropriate referents (Uniform Resource Identifiers or NIL) to mentions of entities, for instance persons

1. http ://pleiades.stoa.org
2. http ://data.bnf.fr

or places, identified in textual documents. This kind of semantic annotation enables later on, by means of such URIs, the querying within distributed collections of documents, and also the automatic retrieval of complementary information about the entities from different sources for aggregating data and building mashups (e.g. thematic maps, historical timelines, etc.).

NEL is typically composed of two steps, namely candidate selection and candidate ranking (Mihalcea et Csomai (2007)), both rely strongly on Linked Data as external Knowledge Base (KB) in the case of unsupervised approaches (Han et al. (2011); Alhelbawy et Gaizauskas (2014); Usbeck et al. (2014); Brando et al. (2015)). The underlying poor quality of LD sets can therefore impact the reliability of a given use case (Paulheim et Bizer (2014)). In the context of NEL, the scope of the data sets chosen as KB can have an important influence on the performances, as texts often concern specific domains of knowledge (e.g. French literature, genetics, zoology), different time spans (e.g. Antiquity, the 19th Century) or particular geographical areas (e.g. France, Asia or even the entire World). There are also issues related to the degree of completeness of the data sets with respect to the different spellings of entity labels (e.g. multilingual and ancient place names) or any other descriptive information available about the entities (e.g. date of birth of persons). The underlying categories that are declared in the ontologies and are used to type entities, for instance, in DBpedia, types correspond to the class hierarchy defined in the DBpedia ontology [3] and to the relational and conceptual categories derived from the Wikipedia category model (e.g. SymbolistPoets [4] or WritersFromParis [5]) for Yago2 by Suchanek et al. (2007), may sometimes have limited exhaustivity or granularity. And last but not least, the existence of SameAs or similar equivalence predicates is crucial for interconnecting descriptions of entities represented in other data sets with different points-of-view.

The present work is outgoing research whose preliminary results are expected to help improving the performance of a NLP pipeline for the massive semantic annotation and enrichment of domain-specific French Literary digital editions [6]. As LD sets vary in quality and are usually generalistic (e.g. DBpedia, Yago2) with few exceptions such as Geonames or Pleaides, we thus aim at quantitatively identifying weaknesses in existing Linked Data for NEL processes according to our needs and considering state-of-art LD quality assessment metrics, in order to further design appropriate solutions in future work. There exists a considerable amount of research works on LD quality, see the extensive literary review by Zaveri et al. (2015). However, they have not yet focused on assessing the value of LD sources for its use in NLP. It is worth mentioning the present efforts for the integration of the two fields [7]. The remainder of this article is organized as follows, section 2 describes LD quality aspects which should be considered for improving NEL in domain-specific contexts. In section 3, we define these quality aspects in the form of metrics identified as useful for NEL, then propose an evaluation procedure for each metric and for both French DBpedia and BnF data sets in order to discuss the opportunity of using these data sets in some domain specific NEL application. Finally, we conclude and draw suggestions for future work.

3. http ://mappings.dbpedia.org/server/ontology/classes
4. http ://dbpedia.org/class/yago/SymbolistPoets
5. http ://dbpedia.org/class/yago/WritersFromParis
6. http ://obvil.paris-sorbonne.fr/bibliotheque
7. See the Linguistic Linked Open Data cloud here : http ://ldl2015.linguistic-lod.org

2 LD quality aspects related to NEL

Maximizing NEL performances implies rating LD sources by means of metrics designed for evaluating precise quality aspects identified as crucial for NEL processes. NEL is usually decomposed in two main steps, candidate selection (including extraction and search) and candidate ranking (Mihalcea et Csomai (2007); Hachey et al. (2013)). For every step, it is possible to measure the local performance using measures such as candidate recall, candidate precision, disambiguation accuracy, among others (see formal definitions in Hachey et al. (2013)). We describe in the following subsections both NEL steps and the challenges concerning LD along with the corresponding quality aspects that, according to our experience, should be evaluated to maximize the performance of every NEL step.

2.1 LD sources quality aspects : candidate selection step

The first phase of a NEL process consists in building a dictionary (or index) of potential candidates, which are extracted from an external KB (see details about one of these strategies in Frontini et al. (2015)). This is typically done only once. This dictionary is usually built by directly querying Sparql endpoints or by pre-processing RDF dumps that are available online. It typically contains alternative surface forms of entities. For instance, the man known as Molière is officially named Jean-Baptiste Poquelin. In this case both real name and pseudonym need to be included in the dictionary. It is thus crucial to rely on as much entity label spellings as possible, for instance, through the skos :altLabel property. This initial dictionary also needs to be in line with the domain of the analyzed text [8].

Besides, for domain-specific NEL, it is also common to reduce the scope of the potential candidates to be included in the dictionaries in order to increase precision. In other words, it is crucial not to include too many entities in the initial dictionary. For instance, for NEL of persons, instead of adding the whole set of persons available in the KB, data are typically filtered using Sparql queries by specifying restrictions depending on the domain, typically temporal (e.g. people born between 1800-1900), geographical (e.g. people who have had any activity in a European country) and thematic (e.g. people who have been involved in some creative work). The level of detail of data filters defined through Sparql queries highly depends on the richness of the model underlying the LD sets involved (i.e. the properties such as "date of birth" or "has worked in" provide simple and effective filters) and the data compliance to the model (i.e. whether the properties have been actually asserted). Some of these challenges in NEL have been highlighted by Rao et al. (2011).

Afterwards, the NEL algorithm searches, for a given text and mention such as "Hugo", a set of entity candidates, for instance the famous French writer Victor Hugo, the well-known translator of the Shakespearean work, François-Victor Hugo, and so on. Candidates are extracted from the previously mentioned dictionary of entities using string matching or similarity heuristics.

The previous processes are closely related, though we described below the quality aspects which concern mostly the extraction of candidates and less their search.

8. In some cases string similarity algorithms may overcome minor spelling variations, but not major ones, which require proper information on aliases.

2.1.1 Domain scope

The first quality aspect is domain scope, i.e. to what extent the data set can be filtered so as to best fit the universe of discourse of the text to be processed before even running the NEL algorithm. In other words, it is the possibility of selecting or filtering the most appropriate entities of the domain, excluding those that are *a priori* irrelevant. This is closely related to what Zaveri et al. (2015) call relevancy of a LD set which is based on two criteria : how the information provided by the dataset fits the use case needs, and the exhausitvity and level of detail of the dataset with regard to the use case. In order to properly filter entities within the domain scope, they need to instantiate the appropriate properties and concepts. For instance, if the text to be processed deals with French writers of the 19th century, the property concerning the date of birth of authors should be systematically instantiated. Otherwise, it would not be possible to properly select the writers who were born before 1900. Likewise, entities lacking of typing statements (e.g. rdf :type or dcterms :subject) are also source of errors, for example the entity Berlin [9] in French DBpedia misses, at the moment of writing, its inherent types PopulatedPlace and SpatialThing. Besides, entities in the data set which do no not instantiate fine-grained concepts (e.g. Person instead of Writer) may also be difficult to select without introducing noise into the dictionary. However, some concepts defined in the KB ontology may not be instantiated at all, or may be instantiated by an insufficient number of resources. In such cases, it is impossible to take advantage of the fine-grained filtering opportunities they could have offered. For instance, the concept FictionalCharacter defined in the DBpedia ontology has not yet been well-exploited, whereas we do find in literary texts many references to fictional characters.

2.1.2 Population completeness

The population completeness is another aspect closely related to the previous one and it represents the percentage of real World entities of a particular type that are represented in the data set. For the context of old French Literary texts, it is common to find for instance mentions of writers who are relatively well-known but they have not yet been listed in current knowledge bases. Indeed, their underrepresentation strongly impacts the candidate recall of NEL.

2.1.3 Alternative labeling richness

The third aspect is alternative labeling richness of entities. This implies the availability of both standard and rejected forms as well as multilingual spellings for each entity. This is close to what Zaveri et al. (2015) call versatility of a dataset and more precisely checks whether data is available in different languages. In our use case, we are also interested in checking whether pseudonyms or alternative names and spellings are available for each language (especially in French). Indeed, the quantity and the diversity of alternative spellings of named entities in the dictionary improves the selection of the right candidates for a mention.

9. http ://fr.dbpedia.org/page/Berlin

2.2 LD sources quality aspects : candidate ranking step

The second phase of NEL consists in choosing the candidate with the highest score for each mention. Due to the lack of annotated French Literary texts we privilege at the moment unsupervised approaches. In this context, RDF graph-based approaches have proven their value (Han et al. (2011); Usbeck et al. (2014); Alhelbawy et Gaizauskas (2014); Brando et al. (2015)). They solely depend on the triplets contained in the KB and are based on the notion of graph centrality or ranking. Scores are commonly measured using Page-Rank or Degree Centrality algorithms. They are computed based on the edge structure underlying the RDF graph where nodes are resources representing entity candidates and edges are predicates linking them. Usually the presence of edges (e.g. dbpedia-owl :influences) or intermediary nodes (e.g. rdf :type) which are common to several nodes (i.e. candidates) influences positively the disambiguation accuracy score. In other words, a graph with a high number of connections between nodes will be more likely to assign correct referents to mentions. Thus it is crucial to retrieve as many RDF statements and shared categories present in the form of intermediary nodes. It is also equally important to retrieve as many triplets as possible for every candidate from different sources.

2.2.1 Granularity of categories in the ontologies

The first quality aspect we identify is the granularity (or exhaustivity) of the underlying categories in the ontologies that are used to type entities via typing predicates. This information constitutes knowledge about context of candidates and contributes to provide their optimal ranking. Indeed, candidates initially retrieved with general criteria are more likely to be good candidates if they also share more specific properties such as belonging to more specific sub-categories. Paulheim et Bizer (2014) studied a similar but not equivalent aspect which is defined as ontology deepness. In our case, it is not only about the amount of hierarchy levels in the ontology but also about how precisely instances are typed by means of fine-grained categories (i.e. top-level vs. leaf-level categories). In general, top-level categories such as Person comprise a large amount of entities, on the other hand, leaf-level categories group fewer entities which share common and specific features such as the Yago category SymbolistPoets that comprises a subset of authors belonging to the Symbolism movement [10]. Having only top-level categories would risk that all candidates share exactly the same categories, such as Person ; the presence of leaf-level categories instead increases the likelihood that some candidates (but not other) share some significant and specific features. So to have sufficient context information about candidates, it is important that the data set should have at least three levels in the hierarchy and also check how many candidates have been typed with leaf-level categories.

2.2.2 Presence of Intra-type relations

Similarly to the previous quality aspect, having as much information as possible about the context of candidates is crucial to propose an optimal ranking of candidates per mention. The data set needs to contain enough RDF statements concerning and shared by at least two candidates (whether in the subject or the object side) of the same type. We name this aspect as the presence of intra-type relations which represents the existence of predicates among entities

10. http ://www.poetes.com/symbolisme

of the same type [11], for instance, in the context of 19th Century French writers, an intra-type relation would be a triplet concerning two French authors via the predicate dbp :isInfluencedBy. In our case, evaluating the presence of intra-type relations focuses on the assertion of object properties between resources of the same category. To motive the use of this metric in a larger context, it is noteworthy to mention that this metric is not only important for NEL but for a whole family of tasks that we can dub as, graph-based disambiguation tasks. In fact, graph-based algorithms for NEL are based on firstly developed Word-Sense Disambiguation approaches (Sinha et Mihalcea (2007)), here too the presence of horizontal, inter-leaves relations can be beneficial w.r.t having only vertical paths connecting leaves. For instance, for disambiguating the word "bank" in the context of "money" and "loan", it would be more beneficial to know an extensive sense of the word "bank" that we denoted bank(1) which is the "institution" who keeps "money" and provide "loans", than only knowing that the bank(1) is an "institution" or a geographical feature (second sense bank(2)). In the field of NLP and computational semantics such problems are quite frequent. Many theories of lexicon, such as Generative Lexicon (Pustejovsky (1991)), put a strong emphasis on such type of horizontal relations between senses to explain the way to interpret the meaning of words. At the same time, it is easily imaginable that such class of disambiguation problems (for which the metric is crucial) may extend beyond the field of language technologies ; this could be the object of further investigation.

2.2.3 Interlinking completeness

Finally, interlinking completeness is an intrinsic LD quality aspect that has been studied by many research works, see an extensive review in Zaveri et al. (2015). It refers to the degree to which entities in the data set are interlinked with other data sources, it can be translated into the existence of several owl :sameAs predicates (or skos :exactMatch). In NEL, it is crucial for accessing multiple representations of a same entity and gathering as much data as possible about it.

3 Evaluation of LD sources against NEL quality aspects

In this section, we firstly define the metrics for evaluating the aforementioned quality aspects and position them in LD quality assessment research, in particular the extensive literary review by Zaveri et al. (2015). We then propose an evaluation procedure for each metric and for several LD data sets in order to measure to what extent the given data sets are well-suited to be used as KB for unsupervised graph-based NEL in domain-specific contexts. Here, we focus on person entities, in particular 19th Century French authors, but the adaptation of the present methodology to deal with other kind of entities such as place, is straightforward. We choose a broad-coverage but rich data set such as French DBpedia [12] and a specialized one such as the BnF [13] data set. More generally speaking, the criteria for selecting LD sets are :

11. It could also be relevant to include RDF statements involving candidates of different types (e.g. persons and places), it depends how the disambiguation approach works, some disambiguate one class at a time and others do not make any distinction.
12. http ://fr.dbpedia.org/sparql
13. http ://data.bnf.fr/sparql

they need to be online and accessible through an Sparql endpoint [14] and they should be related to the domain at stake. Finally, we briefly describe implementation aspects, and also present and discuss evaluation results.

3.1 Quality metrics and evaluation procedures

Definition 1 : Domain scope (DS), the idea here is to obtain the set of entities from a given LD set which may potentially be concerned by the domain at stake and that would be included in the dictionary of entities. In most cases, scope of a data set describing real World entities (e.g. persons, places) can be delimited by three dimension, namely, temporal, spatial and thematic. The filtering of the data set can thus be done in terms of properties and concepts which are delimiters related to one or more of the previous dimensions. These are chosen according to the types of named entities extracted from the texts (e.g. persons). To evaluate the domain-scope of a data set, we choose then the set of properties or concepts that may be consistent with named entities extracted from texts and then check how many resources actually instantiate these concepts and assert these properties. In other words, we count the number of entities within the domain scope, i.e. those fulfilling these conditions.

In formal terms, the domain scope of a LD set DS_{LD} is the size of the set of entities of the set E_{LD} which fulfill every given filter statement composed by a given predicate P and the expected asserted value V, that is,

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

The domain filters we chose are detailed below per data set.

— French DBpedia : Persons who have written something and are labeled as French writers from the 19th Century (temporal and thematic filtering), see Sparql query in Annex.

— BnF : Persons who have written work in French and were born before 1900 (temporal and thematic filtering), also see Sparql query in Annex.

This measure is very close to the metric coverage described by Zaveri et al. (2015) and denoted by R2 which consists in computing the number of entities available in a dataset with regard to the use case. For the time being, the evaluation of relevancy in the sense of R1 metric (Zaveri et al. (2015)), i.e. relevant terms within meta-information attributes, is left to the expert who defines the dataset filtering criteria, and should be further investigated.

Definition 2 : Population completeness (PC), for evaluating the degree of completeness of entities in a data set with respect to the real World entities, equivalent to the metric CM3 described in Zaveri et al. (2015). We would need external and exhaustive information about the latter. As it is not possible to make such an absolute evaluation, we offer instead a relative comparison between the two LD data sets, by means of a manually annotated test set.

In general terms, given an annotated text whose author mentions are assigned, w.r.t ground truth, to URIs of the corresponding entities of a LD set, this set of entities is denoted by E_{text}, given also the set of entities (of the same LD set) which belong to the domain scope E_{DS}, the population completeness of a LD set PC_{LD} is then the intersection set of entities in the domain scope and the entities mentioned in the gold standard, more formally,

14. Efforts have been made for providing information about connectivity of sparql endpoints, the following monitoring tool helps to check their status, http ://sparqles.ai.wu.ac.at.

$$PC_{LD} = E_{text} \cap E_{DS}$$

For French DBpedia and BnF, we use the same domain scope and compare to authors' annotated by humans in a French Literary text, "Refléxions sur la littérature" by Albert Thibaudet (i.e. the gold standard, henceforth called Thibaudet) [15].

Definition 3 : Alternative labeling richness (LR) is the count of labeling predicates present in the set of entities within the domain scope. We distinguish multilingual spellings and compute their distribution between French and other languages. This metric corresponds to versatility of a data set as defined by Zaveri et al. (2015) and it is denoted by V2. Besides, we consider aliases (or pseudonyms) spellings, among others. We also provide the following values for this metric : average number of labels (including all types of labels) per entity, number of labels of the entity having the greatest amount, average number of labels in French.

For each data set, we made the choices listed below.

— French DBpedia : Choice of predicate rdfs :label.

— BnF : Choice of predicates skos :prefLabel and skos :altLabel

Definition 4 : Granularity (G) of underlying categories can be assessed by computing the number of sub-categories of the top-level category used for querying the KB. In formal terms, the granularity of a linked data set G_{LD} is the size of the set of categories C instantiated by the entities within the domain scope E_{DS} via a given typing predicates, that is,

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

Similarly, the distribution of resources within these sub-categories is an interesting information for evaluating to what degree these sub-categories can be used for desambiguating candidates, in more formal terms, it is the size of the set of entities within the domain scope E_{DS} which instantiate any category via a given typing predicate.

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

For each data set, we made the choices listed below.

— French DBpedia : Choice of categories in relation with the predicate skos :broader.

— BnF : Choice of categories concerning the thematic activities described by the predicate rdagroup2elements :fieldOfActivityOfThePerson.

Definition 5 : Presence of intra-type relations (PR) is simply the average number of predicates relating every potential entity candidate (of the same type) within the domain scope. This is similar to what Zaveri et al. (2015) define as property completeness and is denoted by CM2 which measures the missing values for a specific property. Here, for both data sets, we chose only object properties and count per entity, the average of relations with other entities. More formally, the presence of intra-type relations in a LD set PR_{LD} is the average number of entities of the domain scope E_{DS} that assert a given object property Op, where E_{DS} and O constitute different entities and have the same type T.

15. Available from the Labex Obvil digital library,
here : http ://obvil.paris-sorbonne.fr/corpus/critique/thibaudet_reflexions.xml

$$PR_{LD} = | \{ E_{DS} : (E_{DS}\ Op\ O) \wedge E_{DS}\ != O \wedge (E_{DS}\ \text{a}\ T) \wedge (O\ \text{a}\ T) \} | \div | E_{DS} |$$

Definition 6 : Interlinking completeness (IC) is similarly the average number of equivalence predicates for all potential candidates in the domain scope. We count the average of equivalence links per entity. This measure is straightforward and provides an easy-to-interpret score. It is also rather close to the metric CM4 presented by Zaveri et al. (2015) based on the percentage of instances that are interlinked in a dataset. Computing the average of equivalence links per resource provides more information about the chances of finding complementary information in other datasets. In more formal terms, interlinking completeness of a LD set IC_{LD} is the average number of entities of the domain scope E_{DS} which assert an interlinking property Ip, that is,

$$IC_{LD} = | \{ E_{DS} : (E_{DS}\ Ip\ O) \} | \div | E_{DS} |$$

For each data set, we choose the following criteria.
— French DBpedia : Choice of only owl :sameAs predicate.
— BnF : Choice of only owl :sameAs and skos :exactMatch predicates.

3.2 Implementation and evaluation results

For implementing these procedures, we used Sparql queries to compute every metric, similar to the idea of unit-case testing in software engineering for finding errors in tools (in our case, in data) as proposed by Kontokostas et al. (2014). For information, a sample of the produced Sparql queries, in particular those concerning the selection of the domain scope, is annexed to this paper. We then evaluate the corresponding procedures and the obtained results are presented in Table 1.

It is striking to observe the difference between the number of entities in BnF in the domain scope, 51 673, and those selected in French DBpedia, 1384. As expected, BnF would better cover our domain of interest. Also, we quickly see the large amount of authors annotated in Thibaudet's text that are mostly present in BnF. This is also expected as the BnF is used as bibliothecary resource by scholars in the Humanities. We also notice that in general the presence of labels in French is more important in BnF. Clearly the candidate selection phase in NEL would benefit from the BnF data set for this domain. We notice unfortunately that in both data sets, there are many authors missing domain-specific typing information, only 31,6% authors in French DBpedia and 21.8% authors in BnF instantiate categories that may be useful for filtering the scope, at least in the case of BnF as there exist 59 different thematic activities in the ontology. We also observe the absence of relations between authors in BnF which difficult the process of graph-based candidate ranking in NEL, fortunately French DBpedia defines at least some relations which help to compensate. There exist comparable equivalence links in both data sets which is important to gather more equivalent resources out there thus to build a richer graph. This is very important to overcome at certain extent the weaknesses of both data sets in terms of intra-type relations and shared domain-specific categories, thereby to provide an optimal ranking of candidates. These preliminary results allow us to confirm some of our a priori beliefs about the content of both data sets and their usability for domain-specific NEL.

	French DBpedia	BnF
DS_{LD} - # of entities in the domain scope	1 384	51 673
PC_{LD} - # of entities in the intersection of the domain scope and the annotated authors in Thibaudet	207	721
LR - Avg number of all kinds of labels per author	3,1	2,5
LR - # of labels of the author having the greatest amount	19	50
LR - Avg number of labels in French	1	2,47
$G1_{LD}$ - # of different sub-categories instantiated in domain scope	4	59
$G2_{LD}$ - # of authors in domain scope related to any of these sub-categories	469	11 237
$G2_{LD}$ - % of authors in domain scope relating to any of these sub-categories	31.6%	21.8%
PR_{LD} - Avg of relations with other authors per author	0,5	0
IC_{LD} - Avg of equivalence links per author	4,3	5,16

TAB. 1 – *Results of the per metric evaluation of French DBpedia and BnF for NEL of 19th Century French authors.*

4 Conclusions and Future work

In this paper, we presented outgoing research concerning the use of the French DBpedia and the BnF linked data sets in the NLP task of linking of authors in old French Literary texts. The aim of this paper was to discuss some first LD quality metrics that may be relevant in this context and to foster discussion on these issues. Preliminary results showed some of their weaknesses and will help us to further design appropriate solutions to improve our NLP pipeline for the automatic annotation and enrichment of digital editions. We intend to complete our experiment by measuring the performances of the unsupervised graph-based NEL tool REDEN [16] by Brando et al. (2015), in terms of disambiguation accuracy as well as candidate recall and precision, in order to check whether the proposed quality metrics are actually appropriate. In other words, we need to check if the results would be consistent to the preliminary results obtained here so as to verify that the proposed metrics are the most relevant to our needs. Further work is absolutely required to normalize the proposed quality metrics. It is also important to perform new tests on more data sets such as Wikidata or Yago2. Finally, we believe this work

16. REDEN is open source and the code is available here : https ://github.com/cvbrandoe/REDEN

would provide an interesting feedback to the Linked Data research community about how NLP researchers are using LD and the ways these sets can be improved.

Références

Alhelbawy, A. et R. Gaizauskas (2014). Graph ranking for collective named entity disambiguation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Baltimore, Maryland, pp. 75–80. Association for Computational Linguistics.
English

Brando, C., F. Frontini, et J.-G. Ganascia (2015). Disambiguation of named entities in cultural heritage texts using linked data sets. In *New Trends in Databases and Information Systems*, Volume 539 of *Communications in Computer and Information Science*, pp. 505–514. Springer.

Frontini, F., C. Brando, et J.-G. Ganascia (2015). Domain-adapted named-entity linker using linked data. In *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems in Conjunction with the 1st Workshop on Natural Language Applications : completing the puzzle*.

Hachey, B., W. Radford, J. Nothman, M. Honnibal, et J. R. Curran (2013). Evaluating entity linking with wikipedia. *Artif. Intell. 194*, 130–150.

Han, X., L. Sun, et J. Zhao (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*, SIGIR '11, New York, NY, USA, pp. 765–774. ACM.

Kontokostas, D., P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, et A. Zaveri (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, New York, NY, USA, pp. 747–758. ACM.

Mihalcea, R. et A. Csomai (2007). Wikify ! : Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, pp. 233–242. ACM.

Paulheim, H. et C. Bizer (2014). Improving the quality of linked data using statistical distributions. *Int. J. Semant. Web Inf. Syst. 10*(2), 63–86.

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

Rao, D., P. Mcnamee, et M. Dredze (2011). Entity linking : Finding extracted entities in a knowledge base. *Multilingual Information Extraction and Summarization*.

Sinha, R. S. et R. Mihalcea (2007). Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC*, Volume 7, pp. 363–369.

Suchanek, F. M., G. Kasneci, et G. Weikum (2007). Yago : A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, pp. 697–706. ACM.
English

Usbeck, R., A.-C. N. Ngomo, M. Röder, D. Gerber, S. Coelho, S. Auer, et A. Both (2014). AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, et C. Goble (Eds.), *The Semantic Web – ISWC 2014*, Volume 8796 of *Lecture Notes in Computer Science*, pp. 457–471. Springer International Publishing.

Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, et S. Auer (2015). Quality assessment for linked data : A survey. *Semantic Web Journal*.

Annex : Sparql Queries related to domain scope selection

In this annex, we list the qparql queries which served to select the entities concerning the domain of interest, here 19th French authors, firstly French DBpedia, secondly BnF.

```
PREFIX prop-fr: <http://fr.dbpedia.org/property/>
PREFIX dcterms: <http://purl.org/dc/terms/>
select distinct ?ecriv where {
    ?ecriv dcterms:subject ?c .
    ?c rdfs:label ?lp .
    ?ecriv rdfs:label ?l .
    FILTER regex(?lp, ".crivain fran.ais du XIXe si.cle") .
    OPTIONAL { ?ecriv prop-fr:nomDeNaissance ?ndn } .
}
```

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX bnf-onto: <http://data.bnf.fr/ontology/bnf-onto/>
SELECT distinct ?auteur  WHERE {
    ?auteur rdf:type foaf:Person .
    ?auteur foaf:familyName ?nom .
    ?auteur rdagroup2elements:languageOfThePerson ?langue .
    ?auteur bnf-onto:firstYear ?birthdate .
    FILTER (?birthdate < 1900).
    FILTER regex (str(?langue),
    "http://id.loc.gov/vocabulary/iso639-2/fre").
}
```

Summary

See first page.