Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales

Marc Bertin, Tomi Kauppinen, Iana Atanassova

Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales

Iana Atanassova*, Marc Bertin**, Tomi Kauppinen***

*Centre Tesnière, Université de Franche-Comté, Besançon, France
iana.atanassova@univ-fcomte.fr
**CIRST, Université de Québec à Montréal, Montréal, Canada
bertin.marc@gmail.com
***Department of Computer Science, Aalto University School of Science, Finlande
tomi.kauppinen@aalto.fi

Résumé. Nous nous intéressons à la représentation des informations géographiques extraites d'articles scientifiques. En utilisant des outils de Traitement Automatique des Langues et de géo-codage, nous avons traité la revue PLOS Neglected Tropical Diseases afin de produire des données spatiales liées aux articles sous forme de Linked Data. Les résultats montrent une exploitation spatiale et temporelle des données représentant les zones du monde concernées par les maladies tropicales à différentes échelles.

1 Introduction

Ces dernières années, de nombreuses recherches sont menées sur l'exploitation de sources textuelles en complément des systèmes de surveillance épidémiologique (Arsevska et al., 2014). L'extraction d'informations spatiales d'un document reste une tâche non triviale comme le souligne Tahrat et al. (2013). Nous proposons dans cet article une approche pour étudier les maladies tropicales en fonction des informations spatiales extraites d'articles scientifiques. En effet, certaines revues thématiques permettent d'envisager des traitements géographiques à partir d'une étude de corpus des métadonnées liées à l'article. Dans la majorité des cas, les mots-clés ne donnent pas d'informations sur les lieux ou les périodes étudiés. L'enrichissement des méta-données classiques avec des données spatiales et temporelles provenant des textes permettra d'analyser les études scientifiques du point de vue spatio-temporel (Kauppinen et al., 2013). Une telle approche permettrait de répondre à des questions comme *« Quelles régions ont été étudiées en relation avec la Dengue ? »* ou *« Quels lieux ont des co-occurrences dans des études scientifiques dans tel ou tel domaine ? »*. Une problématique plus générale consiste à rendre compte de la nature d'un objet via son association avec des attributs géographiques, tels que des noms de villes, pays, régions et géo-coordonnées. Par exemple, la présence de noms de lieux explicites et de géo-coordonnées associées à des objets, tels que des pages web (Wang et al., 2005b; Borges et al., 2007; Inoue et al., 2002), rendent possible des analyses de ces objets en tant que phénomènes spatiaux. De nombreux travaux proposent des techniques d'enrichissement de ressources par l'établissement de relations avec des informations spatiales

dans des contextes différents (Jones et al. (2002); Wang et al. (2005a); Bucher et al. (2005); Purves et al. (2007); Markowetz et al. (2005)).

Dans cet article nous proposons une méthodologie permettant d'identifier et d'extraire des localisations à partir des publications scientifiques. Notre hypothèse de travail est que l'extraction, la catégorisation et l'affectation de ces données spatiales permettront, à travers un enrichissement des méta-données, d'exprimer les propriétés spatiales des études scientifiques, en établissant des liens avec des localisations géographiques. Notre approche utilise des outils issus du Traitement Automatique de Langues afin de traiter les informations spatiales dans des documents en plein texte.

Notre principale contribution dans cet article est de proposer une lecture des localisations géographiques à travers le filtre des articles scientifiques. Nous illustrons différentes méthodes de visualisations faisant partie des applications possibles. Ces résultats sont destinés à faciliter des décisions pour cibler de nouvelles études, à la veille scientifique et au développement de systèmes de recherche d'information orientés autour des données spatiales.

2 Méthodologie

L'extraction des termes de localisation à partir de corpus d'articles scientifiques permet de représenter la dimension spatiale des études afin de répondre à la question *« Quelles sont les régions / les localisations géographiques / ... qui sont liées à cette étude ? »*. Pour cela, nous cherchons à lier les articles, considérés comme des objets, à des informations spatiales. Dans cette approche, nous excluons les méta-données, telles que les affiliations et les adresses des auteurs. Après l'identification des termes de localisation dans des textes, notre méthode s'appuie sur une désambiguïsation des termes extraits et l'identification des géo-coordonnées afin de produire des visualisations.

2.1 Corpus

Comme corpus d'expérimentation, nous avons utilisé les articles en plein texte du journal *PLoS Neglected Tropical Diseases (PLOS NTDs)*, qui est publié par la *Public Library of Science* [1] et disponible en libre accès. PLOS NTDs publie des articles de recherche examinés par des pairs, dans le domaine des maladies tropicales peu étudiées et traitent de leurs aspects scientifiques, médicaux et sanitaires. Le choix de ce journal a été motivé par sa portée — *"a group of poverty-promoting chronic infectious diseases, which primarily occur in rural areas and poor urban areas of low-income and middle-income countries"* — qui suggère que le fait de rendre explicites les localisations présentes dans les textes en tant que métadonnée apportera des informations précieuses sur les maladies en question.

L'observation montre que les articles de recherche publiés dans PLOS NTDs contiennent un grand nombre d'informations géospatiales. Tous les articles sont disponibles en accès libre en format XML en utilisant le schéma *Journal Article Tag Suite (JATS)* [2]. Nous avons traité l'ensemble complet de 1 872 articles de recherche qui ont été publiés sur une période de 6 ans dans PLOS NTDs, d'août 2007 à août 2013.

1. `http://www.plosntds.org/`
2. Ce standard est une application des normes NISO Z39.96-2012. JATS est une extension de *NLM Archiving and Interchange DTD* (`http://jats.nlm.nih.gov`)

2.2 Extraction des termes de localisation géographique

Afin d'implémenter l'extraction des termes de localisation géographique des textes, nous considérons les étapes suivantes (voir figure 1) :

1. identification des termes de localisation dans des textes ;
2. filtrage des noms de pays ;
3. géocodage des termes identifiés ;
4. analyse des données spatiales à travers des visualisations et analyse des corrélations.

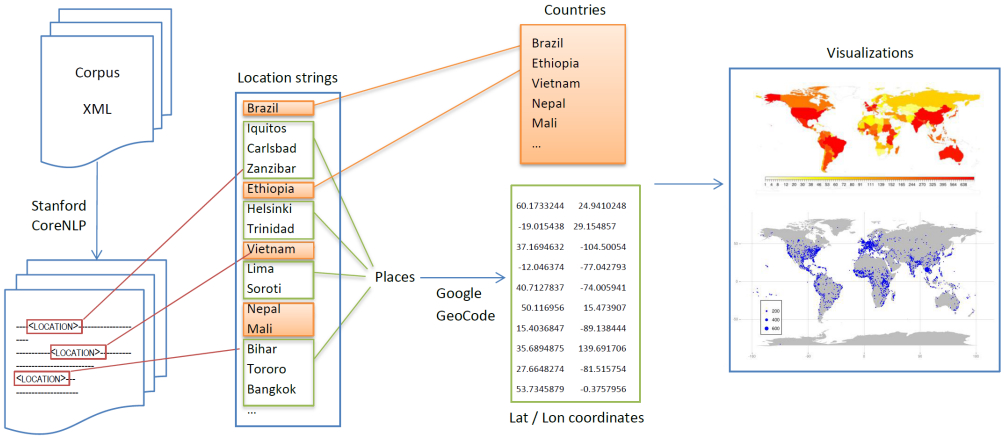

FIG. 1 – *Etapes de l'extraction d'informations spatiales des textes*

Afin d'identifier les localisations, nous nous appuyons sur une Reconnaissance d'Entités Nommées (REN) [3], qui est une technique bien connue en Traitement Automatique des Langues et qui utilise des méthodes d'apprentissage automatique. Le but des systèmes de REN est d'identifier les entités nommées dans des textes et les annoter avec des catégories différentes, telles que *location*, *organization*, *proper name*, *date*. Nous avons réalisé l'extraction des termes de localisation en utilisant l'outil de REN de Stanford CoreNLP [4] (Manning et al., 2014).

Dans notre étude, nous nous intéressons particulièrement aux différents types de localisations : pays, villes ou régions. Par ailleurs, les informations géographiques présentes dans des textes sont souvent incomplètes et ambiguës, ce qui implique des traitements supplémentaires après l'extraction afin d'assurer la qualité des données générées. Par exemple, nos premières expérimentations ont révélé qu'une partie des termes extraits en tant que localisations par le système de REN sont des noms de virus et ne correspondent pas aux localisations géographiques. Des traitements ont été mis en place à des fins de nettoyage :

1. vérification si le terme extrait est le nom d'un pays. Si oui, nous pouvons établir un lien entre ce pays et l'article traité.

3. voir par exemple Cucerzan et Yarowsky (2002); Zhou et Su (2002)
4. Stanford CoreNLP Named Entity Recognizer

2. si le terme extrait n'est pas le nom d'un pays, nous pouvons émettre l'hypothèse qu'il est le nom d'une ville ou une autre localisation plus petite qui pourra être représentée par un point. Dans ce cas, nous utilisons Google GeoCode API [5] afin de retrouver la latitude et longitude.

La distinction entre les pays et les autres localisations (villes, régions) est nécessaire notamment pour des fins de visualisation. En effet, nous supposons que les localisations qui ne sont pas des pays correspondent à des territoires relativement petites et peuvent être représentées par des points sur une carte géographique à l'échelle mondiale. Ainsi, nous avons deux principaux types de données spatiales pour chaque article : pays mentionnés dans l'article et autres termes de localisation dans le texte. Une méthodologie peut être considérée afin de prendre en compte la taille des territoires des localisations identifiées de façon précise pour pouvoir obtenir des visualisations plus fines.

Dans la première étape nous avons utilisé la liste de pays traitée par la norme ISO 3166 [6], qui fournit une correspondance entre les noms des pays en anglais et des codes des pays reconnus internationalement.

La deuxième étape de ce traitement a pour fonction, en plus de l'identification des géo-coordonnées, de filtrer les expressions pour lesquelles Google GeoCode API ne retourne pas de résultats. La table 1 présente le nombre de pays, localisations et géo-coordonnées identifiés dans le corpus. Sur un total de 24 660 occurrences de termes de localisations, autour de 85% (20 990) ont été convertis en géo-coordonnées avec succès. Cette étape permet d'éliminer une partie des entités nommées qui ont été identifiées par CoreNLP et qui ne correspondent pas à des noms de lieux. Cependant, cette méthode ne nous permet pas la distinction entre les différents sens des noms de localisations ambiguës.

TAB. 1 – *Occurrences de pays et termes de localisations dans le corpus*

	Pays	Localisations (CoreNLP)	Localisations avec géo-coordonnées (Google GeoCode API)
Total	24 197	24 660	20 900
Nb moyen par article	12,93	13,17	11,16

Remarquons que certains localisations sont présentes avec une très grande fréquence dans le corpus. De plus, les occurrences de seulement 15 pays et 200 termes de localisations représentent la moitié de toutes les occurrences. La table 2 présente le nombre total de pays et termes de localisation distincts dans le corpus.

TAB. 2 – *Pays et termes de localisation distincts*

Pays	Termes de localisation	Géo-coordonnées
150	4 168	3 249

5. http://maps.googleapis.com/maps/api/geocode/
6. http://www.iso.org/iso/country_codes.htm

2.3 Génération de Linked Data

Pour chaque article, nous avons obtenu une liste de pays et une liste de localisations avec leurs coordonnées latitudinales et longitudinales. Cela nous permet de générer des données sous forme de Linked Data, qui représentent les informations géographiques issues de l'article. Ce format a été choisi afin de pouvoir fournir des données réutilisables pour des applications externes. Elles peuvent être exploitées dans des visualisations, mais également en tant que ressources dans des systèmes de recherche d'informations ou d'extraction de connaissances. Les données sont accessibles à l'adresse suivante : `http://linkedscience.org/data/spatialaboutness/`.

La génération de ces données en tant que ressources indépendantes établit un lien fort entre la géographie et les études de terrain. La publication de ces résultats sous forme de Linked Data permet de mettre en place une base de connaissances incrémentale afin de gérer et partager les données spatiales pour l'utilisation par la communauté via des services SPARQL.

Nous avons utilisé le vocabulaire *Linked Science Core Vocabulary (LSC)* [7], qui a été spécifiquement créé pour représenter des propriétés des recherches scientifiques, y compris les données temporelles et spatiales liées aux études scientifiques. En se basant sur les termes de LSC, nous avons converti les informations géographiques extraites des articles en des triplets RDF.

```
@prefix ns:<http://linkedscience.org/lsc/ns#> .

journaldoi:journal.pntd.0000321
    lsc:isAboutRegion
        aboutloc:Akonolinga, aboutloc:Ayos, aboutloc:Bu,
        aboutloc:Nyong, aboutloc:Thailand;
    a lsc:Research .

journaldoi:journal.pntd.0000355
    lsc:isAboutRegion
     aboutloc:Mahottari, aboutloc:Muzaffarpur, aboutloc:Pune,
     aboutloc:Rajshahi, aboutloc:Vaishali, aboutloc:Varanasi;
    a lsc:Research .
```

FIG. 2 – *Exemples de triplets RDF qui décrivent la dimension spatiale des études scientifiques*

La figure 2 montre deux exemples de triplets RDF représentés dans la syntaxe Turtle [8]. Le champ *journaldoi* permet d'identifier chaque article de façon unique par son *Digital Object Identifier (DOI)*. Le premier exemple exprime le fait que l'article identifié par *journal.pntd.0000321* est une étude scientifique qui concerne les régions d'Akonolinga, Ayos, Bu, Nyong et Thailand.

7. `http://linkedscience.org/lsc/ns/`
8. `http://www.w3.org/TeamSubmission/turtle/`

3 Résultats

Les données spatiales que nous avons obtenues à partir des articles peuvent être exploitées dans plusieurs types d'analyses. Nous proposons différentes visualisations géographiques, ayant pour objectif principal de mettre en évidence les zones géographiques qui concentrent le plus grand nombre de recherches ainsi que les zones peu étudiées. Les visualisations ont été générées en utilisant *R-studio* et la librairie *rworldmap*.

Nous considérons deux représentations principales : une au niveau des pays et une sur une échelle plus fine. De plus, comme le corpus que nous avons traité est homogène et couvre six ans, nous pouvons examiner et visualiser les tendances, en prenant en compte la date de publication des articles.

3.1 Analyse au niveau des pays

La visualisation sur la figure 3 montre une carte thermique des pays mentionnés dans les articles du corpus. Comme le montre cette représentation, certaines régions ont fait l'objet d'un très grand nombre d'études, par exemple l'Afrique du Sud, alors que d'autres régions comme l'Afrique Centrale sont peu étudiées. Cette carte donne ainsi un premier aperçu des régions constituant des centres d'intérêt pour le journal.

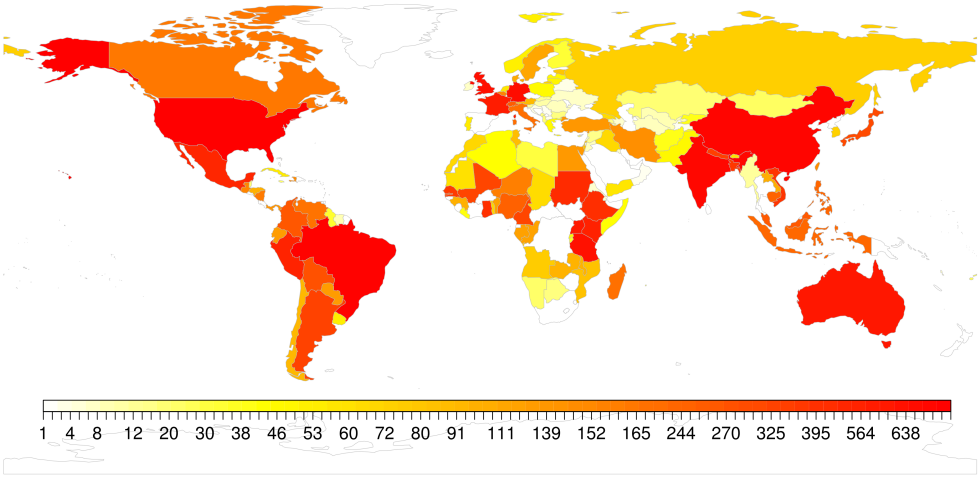

FIG. 3 – *Pays mentionés dans le corpus par nombre d'occurrences.*

Notons que plusieurs pays non-tropicaux sont également présents avec des fréquences élevées, dont les États-Unis, certains pays européens, le Japon, etc. Ces occurrences proviennent bien du corps des textes et non pas des métadonnées des articles. En effet, ces pays sont souvent mentionnés dans le corpus en lien avec des laboratoires, vaccins, établissements hospitaliers, etc.

Si en plus nous prenons en compte les années des publications, les données permettent de détecter et observer les tendances dans la recherche. Par exemple, nous pouvons représenter

sur un graphe le nombre d'occurrences des noms de pays par année, comme le montre la figure 4. En prenant comme exemple le continent d'Afrique, nous avons considéré les cinq pays ayant le plus d'occurrences. Le graphe à gauche présente le pourcentage du nombre d'occurrences de chaque pays par rapport à toutes les occurrences pour une année donnée. Le graphe à droite présente le pourcentage du nombre d'articles qui mentionnent chaque pays par rapport à tous les articles publiés dans la même année. La différence entre ces deux graphes provient du fait que le même article peut contenir de multiples occurrences d'un pays, qui seront alors comptées plusieurs fois pour le graphe à gauche et une seule fois pour le graphe à droite. Comme le corpus couvre la période d'août 2007 à août 2013, les données pour 2007 sont sur seulement 4 mois, et les données pour 2013 sont sur 8 mois.

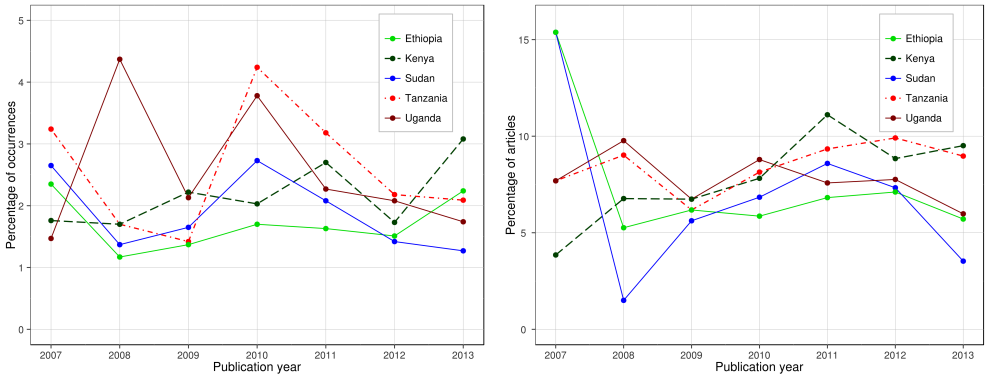

FIG. 4 – *Pays par nombre d'occurrences par année : les 5 pays en Afrique les plus cités*

Ces résultats montrent que les 5 pays le plus fréquemment mentionnés en Afrique se trouvent en forte proximité géographique : Kenya, Tanzanie, Ouganda, Soudan et Éthiopie. Les graphes indiquent un nombre relativement élevé d'occurrences de Tanzanie et Ouganda pour 2009 et 2011, ainsi qu'un intérêt émergeant pour le Kenya jusqu'à 2011. En 2012 et 2013, Kenya et Tanzanie sont cité dans presque 10% des articles, alors que le nombre d'occurrences des autres pays diminue. Globalement, nous pouvons observer une hausse significative du pourcentage d'articles liés à cette région entre 2009 et 2011 : en effet, en 2011 les articles mentionnant ces cinq pays constituent plus de 35% de tous les articles.

Nous avons également examiné les co-occurrences entre les pays les plus cités. La figure 5 montre une partie des corrélations obtenues. Nous pouvons observer les corrélations les plus importantes entre Ouganda et Kenya (0,28) et entre Soudan et Éthiopie (0,26). Cependant, ces valeurs ne sont pas élevées, ce qui signifie que peu de pays apparaissent ensemble dans les articles de façon systématique.

3.2 Analyse utilisant les géo-coordonnées

Pour toutes les localisations qui ont été extraites des textes, nous avons obtenu le nombre d'occurrences et les géo-coordonnées. La carte sur la figure 6 montre toutes les localisations géographiques. Les tailles des points sont relatives aux nombres d'occurrences. La grande

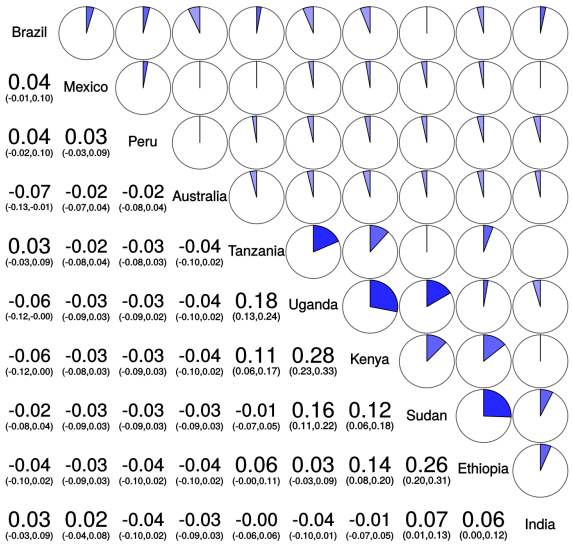

FIG. 5 – *Corrélations entre occurrences de pays*

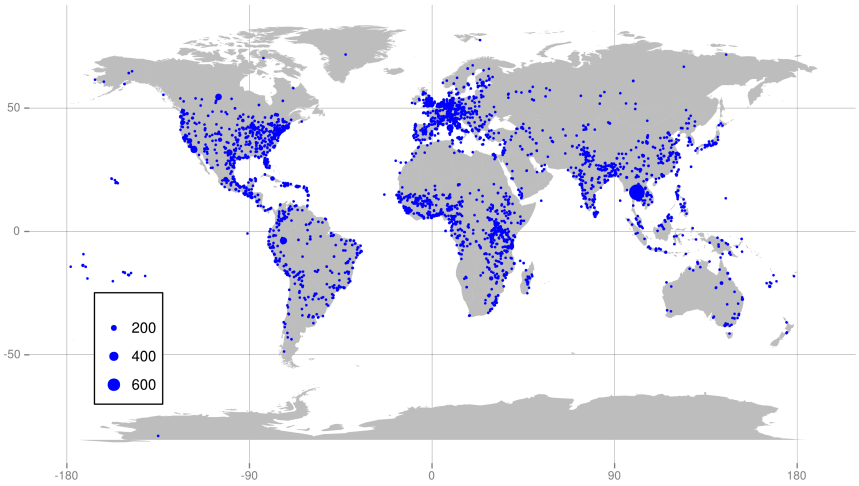

FIG. 6 – *Visualisation de localisations géographiques dans le corpus. Les tailles de points correspondent aux nombres d'occurrences.*

concentration de points dans certaines régions indique leur importance du point de vue de l'étude des maladies tropicales.

Ces données peuvent être exploitées également en prenant en compte les années de publication afin de visualiser l'évolution du nombre de recherches liées à chaque région. La figure 7 montre les localisations géographiques pour des différentes années de publication [9]. L'émergence de nouvelles régions d'intérêt peut être observée. Par exemple, un nombre croissant d'études entre 2008 et 2013 concernent la région autour de Bangkok, Thaïlande.

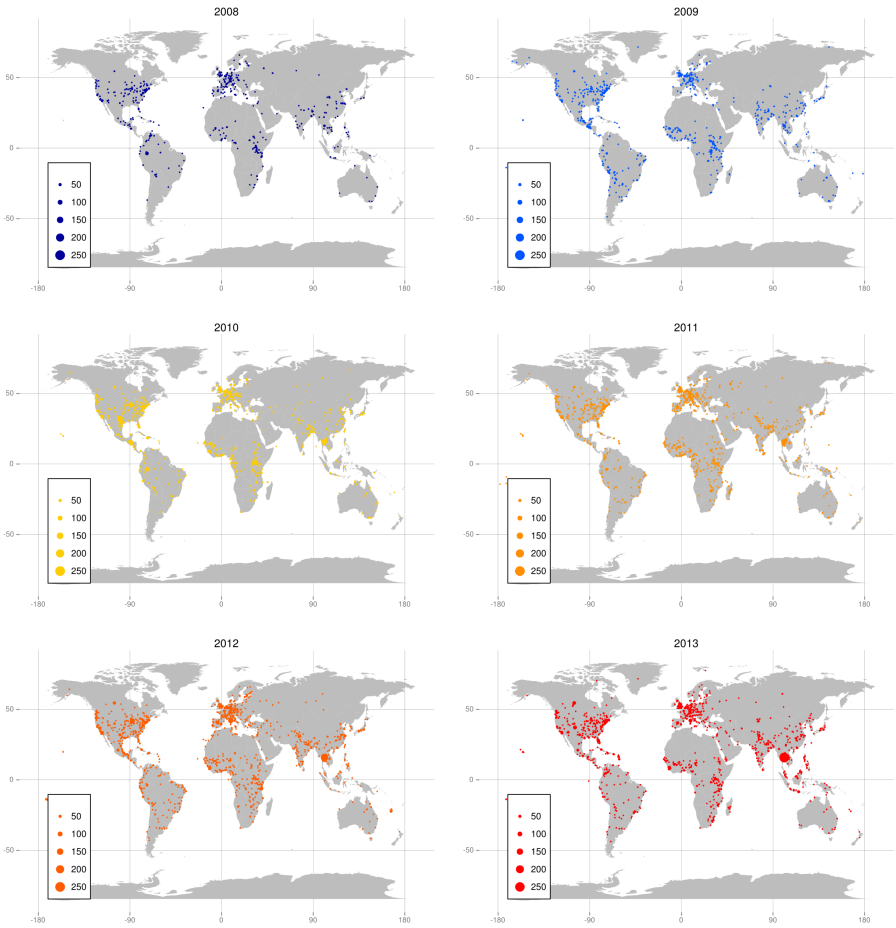

FIG. 7 – *Visualisation de localisations géographiques par année*

9. Une interface en ligne permettant l'exploitation interactive de ces données se trouve à l'adresse suivante : `http://linkedscience.org/demos/spatialaboutness/`

4 Discussion

Cette étude a pour but de montrer la pertinence de l'exploitation de données géo-spatiales dans le cadre de corpus scientifique spécifique à un domaine comme PLOS PNTDs. Les raisons de la présence de localisations géographiques dans des études scientifiques sont variées. Comme le montrent les exemples dans la table 3, les localisations sont liées à des épidémies, des vaccins, provenances de virus ou d'échantillons, adresses de compagnies médicales, etc.

TAB. 3 – *Exemples de phrases contenant des localisations*

Phrase	Type de localisation
RVFV has led to outbreaks in *Egypt* and the *Arabian Peninsula* with the potential to spread to the *United States* and *Europe*.	Epidémies
Briefly, One Step RT-PCR Kit (Qiagen ; *Valencia, CA*) was used for the RT-PCR reactions.	Adresse de compagnie médicale
A separate, internal control reaction for the detection of RNAse P was performed on the clinical samples from *Nicaragua* and *Sri Lanka*.	Provenance des échantillons
In a 2-site ID vaccine trial in *Thailand*, antibody levels varied 2.2 fold between different hospitals.	Campagne d'immunisation

Les limites de cette étude sont de nature linguistique. D'une part, les noms des localisations géographiques peuvent être polysémiques. La désambiguïsation des entités nommés peut être envisagée en faisant appel à des connaissances encyclopédiques (voir Bunescu et Pasca (2006)). Cependant, dans cette étude nous n'avons pas effectué de désambiguïsation et nous nous sommes appuyés sur les premiers résultats de Google GeoCode API. Il est possible que cette limitation ait introduit des erreurs dans les visualisations. D'autre part, l'extraction des localisations par l'outil CoreNLP ne permet pas de prendre en compte certaines expressions linguistiques qui contribuent à préciser un lieu. Or, il pourrait être utile d'exploiter la distinction entre, par exemple "*southern Tanzania*" et "*rural districts in northwest Tanzania*". La prise en compte de telles variations permettra d'obtenir une plus grande précision dans les représentations.

Il en résulte que la catégorisation des localisations pourra fournir des données pour des analyses plus fines avec des applications en recherche d'information et en veille. Par exemple, cela permettra de distinguer entre les localisations qui sont des adresses de laboratoires ou de compagnies et celles qui sont des foyers de contagion.

5 Conclusion

Notre objectif était de rendre compte de la dimension spatiale des études et de fournir les données nécessaires à une agrégation visuelle des résultats. En s'appuyant sur l'outil de Reconnaissance d'Entitées Nommées de Stanford, nous proposons une approche afin d'extraire les localisations géographiques liées aux études de la revue PLOS Negleted Tropical Diseases. Les résultats montrent l'évolution dans l'espace et le temps des zones porteuses de maladies tropicales comme le montrent par exemple les visualisations obtenues pour la Thaïlande.

Notre futur travail s'articulera autour de la catégorisation des localisations extraites à partir des textes. Un autre objectif est la construction d'un service d'agrégation et de partage de données géographiques liées à des études scientifiques, accessibles via SPARQL [10] ou GeoSPARQL [11] sous forme de Linked Data. Cela permettra la ré-utilisation de ces données par des applications et services externes.

Le fait d'établir des liens entre des études scientifiques et des informations géo-spatiales permet de fournir de nouveaux descripteurs des publications pour l'enrichissement de métadonnées. Cette approche permet une nouvelle lecture des articles scientifiques à travers les données géographiques. Il s'agit d'une première étude montrant des applications autour des représentations spatiales liées aux études scientifiques, notamment par des visualisations géographiques. Les résultats soulignent la nécessité d'une étude linguistique des contextes d'entités nommées dans les textes afin de catégoriser les informations géographiques. Cela nous permettra, à terme, de proposer une ontologie montrant les relations entre laboratoires et foyers de contagion.

Remerciement

Nous remercions Benoit Macaluso de l'Observatoire des Sciences et des Technologies (OST) [12], Montréal, Canada, pour le moissonnage du corpus PLOS.

Références

Arsevska, E., M. Roche, R. Lancelot, P. Hendrikx, et B. Dufour (2014). Exploiting textual source information for epidemiosurveillance. *Metadata and Semantics Research*, 359.

Borges, K. A., A. H. Laender, C. B. Medeiros, et C. A. Davis Jr (2007). Discovering geographic locations in web pages using urban addresses. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pp. 31–36. ACM.

Bucher, B., P. Clough, H. Joho, R. Purves, et A. K. Syed (2005). Geographic ir systems : requirements and evaluation. In *Proceedings of the 22nd International Cartographic Conference*, Volume 201, pp. 11–16.

Bunescu, R. C. et M. Pasca (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*, Volume 6, pp. 9–16.

Cucerzan, S. et D. Yarowsky (2002). Language independent ner using a unified model of internal and contextual evidence. In *proceedings of the 6th conference on Natural language learning-Volume 20*, pp. 1–4. Association for Computational Linguistics.

Inoue, Y., R. Lee, H. Takakura, et Y. Kambayashi (2002). Web locality based ranking utilizing location names and link structure. In *Web Information Systems Engineering Workshops, International Conference on*, pp. 56–56. IEEE Computer Society.

Jones, C. B., R. Purves, A. Ruas, M. Sanderson, M. Sester, M. Van Kreveld, et R. Weibel (2002). Spatial information retrieval and geographical ontologies an overview of the spirit

10. http://www.w3.org/TR/rdf-sparql-query/
11. http://www.opengeospatial.org/standards/geosparql
12. http://www.ost.uqam.ca/

project. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 387–388. ACM.

Kauppinen, T., A. Baglatzi, et C. Keßler (2013). Linked Science : Interconnecting Scientific Assets. In T. Critchlow et K. Kleese-Van Dam (Eds.), *Data Intensive Science*. USA : CRC Press.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, et D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pp. 55–60.

Markowetz, A., Y.-Y. Chen, T. Suel, X. Long, et B. Seeger (2005). Design and implementation of a geographic search engine. In *WebDB*, Volume 2005, pp. 19–24.

Purves, R. S., P. Clough, C. B. Jones, A. Arampatzis, B. Bucher, D. Finch, G. Fu, H. Joho, A. K. Syed, S. Vaid, et al. (2007). The design and implementation of spirit : a spatially aware search engine for information retrieval on the internet. *International Journal of Geographical Information Science 21*(7), 717–745.

Tahrat, S., E. Kergosien, S. Bringay, M. Roche, et M. Teisseire (2013). Text2geo : from textual data to geospatial information. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pp. 23. ACM.

Wang, C., X. Xie, L. Wang, Y. Lu, et W.-Y. Ma (2005b). Detecting geographic locations from web resources. In *Proceedings of the 2005 workshop on Geographic information retrieval*, pp. 17–24. ACM.

Wang, L., C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, et Y. Li (2005a). Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 424–431. ACM.

Zhou, G. et J. Su (2002). Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 473–480. Association for Computational Linguistics.

Summary

In this paper we present an approach for the extraction of geographic information from scientific articles in the biomedical domain. The idea is make sense of geographic dimension of articles via geo-spatial visualizations and enhanced information retrieval. We evaluate the approach by using the journal PLOS Neglected Tropical Diseases as a data source. We perform full-text analysis of the articles and produce Linked Data to describe the geographic aboutness of scientific studies. We make use of visualizations to present the results and discuss their implications for the future work.