



Extraction of Relevant Data from Social Media Based on Terminological Resources: *Application to Meningitis Surveillance via Twitter*

Wend-Panga Régis Cédric Bere^{1,5(✉)}, Gaoussou Camara²,
Sadouanouan Malo³, Sylvie Despres⁴, Moussa Lo⁵, and Stanislas Ouaro¹

¹ LAMI, Université Joseph Ki-Zerbo, Ouagadougou, Burkina Faso
cedric.bere@gmail.com,ouaro@yahoo.fr

² EIR-IMTICE, Université Alioune Diop de Bambey, Bambey, Senegal
gaoussou.camara@uadb.edu.sn

³ Université Nazi Boni, Bobo-Dioulasso, Burkina Faso
sadouanouan@yahoo.fr

⁴ LIMICS, Université Paris 13, Paris, France
sylvie.despres@univ-paris13.fr

⁵ LANI, Université Gaston Berger, B.P. 234 Saint-Louis, Senegal
moussa.lo@ugb.edu.sn

Abstract. In this paper, we present a process for collecting and filtering relevant data for epidemiological surveillance of meningitis. We focus on the African meningitis belt stretching from Senegal to Ethiopia. This study aims to fill the data gap for the early detection of epidemics based on the analysis of social media. Our approach is based on previous work that showed that social media analysis contributes significantly to the surveillance of epidemics. It uses IDOMEN (Infectious Disease Ontology for MENingitis) a meningitis domain ontology and a SKOS resource meningVocab (meningitis vocabulary). IDOMEN is an extension of the Infectious Disease Ontology (IDO). The SKOS resource meningVocab is built from a corpus of meningitis tweets from social media. We align the IDOMEN ontology and the SKOS resource meningVocab for collection and filtering tweets containing data relevant to meningitis in a perspective of epidemiological surveillance. Tweets are collected via the Twitter API on the basis of a list of terms related to meningitis. They are then annotated using these two resources and filtered using the rules of the domain (for example, the rules characterizing situations suggestive of bacterial meningitis: *fever* AND *purpura* AND *headache*).

Keywords: Ontology · SKOS Resource · Social media · Epidemic intelligence · Meningitis

1 Introduction

The advent of social media is contributing to a significant increase in the generating of digital data. Using of this data offers new opportunities in several disciplines and provides many applications in various fields such as biology and health.

Infodemiology is a combination of the words: “*information*” and “*epidemiology*”. This term refers to a science based on Information and Communication Technologies (ICT) to achieve population health monitoring and guide public health policies [7]. Infodemiology exploits various data sources: websites, social media such as “micro-blogging” sites (Twitter, Tumblr etc.). These data are collected and analyzed in real time in order to predict and warn about risks of epidemics.

For more than two decades, meningitis epidemics have not been effectively contained. Meningococcal meningitis is an infectious disease that causes inflammation of the meninges surrounding the brain and spinal cord. This disease is caused by a Gram-negative bacterium called *Neisseria meningitidis*. The World Health Organization (WHO) recorded nearly one million suspected cases of meningitis and more than 100,000 deaths reported in the African meningitis belt. In 2017, Nigeria and Niger experienced a major epidemic with more than 18,000 registered cases¹. The detection of the first cases of meningitis is crucial for taking appropriate measures to contain a possible epidemic. This identification requires the collection and rapid integration of data and events related to risk factors for its propagation. However, in sub-saharan African countries, there is usually no real-time data collection tool for disease surveillance systems. Therefore, there are difficulties in early detection of the meningitis epidemic.

The work of [3] has shown that the use of ontologies can be an appropriate tool for epidemiological surveillance. A meningitis domain ontology could be an effective way to accomplish the task of filtering tweets. Indeed, ontologies provide both domain terminology and axioms describing relationships between domain concepts. In this paper, we present a process of collecting and filtering relevant data for the epidemiological surveillance of meningitis in the African meningitis belt stretching from Senegal to Ethiopia. As part of this work, we use the terminology associated with IDOMEN [1] and meninVocab to implement the process of collecting and filtering tweets containing relevant meningitis-related data. IDOMEN being built from the knowledge of medical experts, the use of a controlled SKOS vocabulary aims to capture tweets posted by communities of non-expert users using their own vocabulary that does not necessarily correspond to the technical terms of the medical field of disease. This controlled vocabulary is constructed from terms used in a manually processed sample of tweets.

The purpose of this work is to propose a collection and filtering process for the early detection of meningitis epidemics using real-time data from the Twitter social media. Our strategy consists in collecting and filtering messages from Twitter using termino-ontological resources (IDOMEN, meninVocab). Then we annotate the

¹ WHO: Controlling Meningitis Epidemics in Africa: A Quick Reference Guide for Health Authorities and Caregivers, <http://www.who.int/csr>.

tweets collected. The annotation allow to identify the terms contained in the Tweets and go back through the relationships in the termino-ontological resources. We perform matching of the annotated elements with IDOMEN concepts that appear in the rules of diagnosis of a “*suspected case*” of meningitis.

In the following of this paper, we draw up a state of the art relating to the extraction of data in social networks. In the second section, we describe the termino-ontological resources (*IDOMEN*, *meningVocab*) used in our approach. Then we present the methodology of collection and filtering adopted and based on the use of termino-ontological resources. Finally, we discuss the results of filtering on the social network Twitter.

2 Related Works

Social networks are web-based technology platforms that allow organizations or individuals to produce and share text, image, video, or audio content. These social networks are new modes of electronic communication but the volumetry of the data produced and the diversity of their format make their exploitation complex. Indeed, the data is mostly unstructured, which makes it difficult for them to be used immediately by machines. The reuse of this data requires an extraction guided by needs of use. The purpose of extracting data from social networks is to analyze them in order to predict a situation, to understand user behavior and to explain societal phenomena.

In this part, we draw up a state of the art focused on health surveillance in social networks. Some works has been done on the health surveillance of diseases such as influenza, dengue and other diseases with high epidemic potential. Most of these studies were conducted in countries in Europe, America and Asia.

In the context of influenza surveillance, there are two types of approaches those based on supervised classification [4] and those based on unsupervised classification [2]. It is possible to combine these approaches with message localization techniques [6] associated with the dynamics of disease propagation, but also with the frequency of the keywords related to the disease studied.

In [6], they are interested in monitoring influenza on Twitter. The tweets ranked according to the relevance of the infection. They combines a tweets location to focus the study on 10 countries from 4 continents having English as the official language. The results of this work show strong correlations between the number of tweets related to an influenza infection and the number of influenza cases recorded by the health surveillance services in these 10 countries.

[10] shows that the social network Twitter can serve as a complementary tool for seasonal surveillance of influenza in France. The authors base their approach on the use of keywords relating to the human flu. They use the Twitter streaming API to collect tweets related to their keyword vocabulary. The tweets relating to the animal flu are automatically deleted from the corpus obtained. To filter relevant tweets, they use a machine-learning algorithm based on SVMs (Support Vector Machine). Subsequently, the analysis of these data consists in comparing the data obtained on Twitter with data from two networks of hospital emergency

structures. Indeed, by using Spearman's correlation coefficients and significance tests according to regions in France, [10] manage to establish that there is a link between the weekly number of tweets relating to influenza and the weekly number of influenza cases recorded in these two networks of health structures.

This work shows that the tools offered by automatic language processing (standardization, lemmatization, N-gram, etc.) constitute the common base for all this work to analyze the content of tweets. The approaches presented are generally based on the use of keywords. Other works show the interest of using an ontology. Indeed an ontology provides a more complete terminology of the field than the keywords. Moreover the exploitation of the relations between the concepts and the axioms make it possible to refine the filtering or the classification of the tweets.

The ambition of the ontology is to provide a conceptual model describing the concepts of the domain, the relations that exist between these concepts, while specifying the nature and the mode of use of these relations.

In this paper, we propose to use termino-ontological resources to enrich the annotation process and take into account the particular nature of tweets.

3 Use of the Termino-Ontological Resources

In this section we present the termino-ontological resources (IDOMEN, meningVocab), we use for collecting and filtering the tweets. Termino-ontological resources refer to a model that has both a conceptual component described/represented by an ontology and a lexical or terminological component associated with the ontology. The terminological component is linguistic in nature and therefore in the vocabulary used in the communities that employ them.

These termino-ontological resources will be used to support the semantic annotation of messages from Twitter.

3.1 The IDOMEN Ontology

The purpose of IDOMEN ontology [1] is to allow the annotation of texts to assist in the extraction of data from social media text messages related to meningitis. IDOMEN will also serve as a support for data integration, sharing of knowledge in the field and contributing to effective communication between the actors of the epidemiological surveillance system. IDOMEN is an extension of IDO ontology, IDO is a core ontology that is common to infectious diseases, it does not address specificities related only to a given disease. IDOMEN deals with meningitis, this ontology covers clinical aspects, biological aspects but above all climatic and social economic aspects which are linked to the risk factors of epidemics. The methodology that IDOMEN ontology uses is based on the NeOn methodology [8]. IDOMEN implements the OBO recommendations advocated by the *Open Biomedical Ontologies*, which correspond to proven good practices that promote interoperability, reusability and good ontology construction. The architecture of the IDOMEN ontology is structured in three modules: the

biological perspective, the clinical perspective and the epidemiological and public health perspective. Each 3 modules divided into sub-modules linked together by the semantic relations existing between a concept of a sub-module with a concept of another sub-module.

In this paper, our collection application is mainly based on the clinical perspective (aspects related to the symptoms) Fig. 2 and the epidemiological perspective (aspects related to risk factors, epidemic emergence factors of meningitis) Fig. 1. These figures show the main hierarchies of IDOMEN concepts whose associated labels will be used in this collection. Patients discuss on social media (forums, web site, microblog site,) and share their experience and history of the disease by explaining the different manifestations of the disease. The manifestations of the disease are most often referred to as symptoms: fever, migraine, convulsion, etc. Figure 1 illustrates the emergence risk factors that characterize the climatic environment and promote the spread of meningitis. Figure 2 shows the main clinical factors related to meningitis.

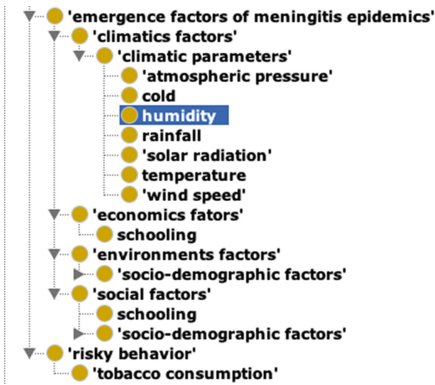


Fig. 1. Concepts related to epidemiological and public health aspects

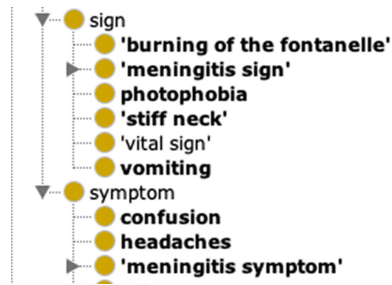


Fig. 2. Concepts related to clinical aspects

3.2 Annotation of the Corpus with IDOMEN Vocabulary

In this part we use only IDOMEN’s vocabulary to annotate tweets to evaluate IDOMEN coverage on a corpus of tweets related to meningitis. Our corpus consists of 1413 tweets related to meningitis. We use GATE (General Architecture for Text Engineering,) [5] to make the annotations of the corpus. Figure 3 illustrates the annotation performed, and Table 1 the result of the annotation process. We note that we have 1780 annotations on the 1413 tweets.

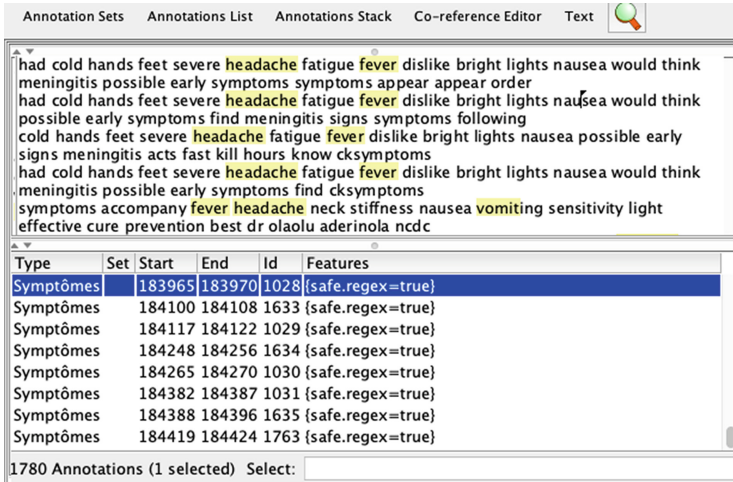


Fig. 3. Annotation of the corpus in GATE with IDOMEN vocabulary

Table 1. Table of annotation results with IDOMEN vocabulary.

#Concepts	rdfs:Label	Number of occurrences
IDOMEN_0000026	Vomiting	97
IDOMEN_0000028	Photophobia	14
IDOMEN_0000029	Stiff neck	215
IDOMEN_0000031	Headache	294
IDOMEN_0000031	Burning of the fontanelle	0
IDOMEN_000032	Confusion	93
IDOMEN_000040	Fever	1031
IDOMEN_0000041	Convulsion	22
IDOMEN_0000044	Purpura	9
IDOMEN_0000045	Neurological signs	5
Total		1780

The results highlight the low coverage of the IDOMEN vocabulary on the tweets corpus. Indeed, manually exploring the tweets it appears that some symptoms common to meningitis are expressed in terms different from those used by the experts domain. The Fig. 3 presents two examples of terms used in Twitter to designate symptoms that are not in the IDOMEN vocabulary. Indeed, Twitter users do not necessarily know the experts terminology and use terms that express the form or nature of the manifestation of this symptom in the language they are familiar with. Thus, to express “photophobia”, users will tend to use

these expressions “eye sensitivity” or “light dislike” earlier than the term “photophobia”. The nature of IDOMEN ontology justifies the findings, IDOMEN is built by the semantic concepts and relationships of the field used by experts: clinicians, epidemiologists of meningitis, biomedical practitioners. In the perspective of keeping this ontology to use later the decision rules determining if a tweet is likely to describe a case of meningitis and which are constructed from IDOMEN’s concepts and relationships, we have built a terminological resource gathering the terms used in Twitter to designate manifestations corresponding to meningel symptoms. In this section, we present the SKOS resource named “meningVocab” and developed from a sample of meningitis-related messages from the Twitter social media (Fig. 4).

chronic conditions know muscle pain fever nausea vomiting symptoms meningitis given
 would like know exactly symptoms occur
 headache nausea meningitis guess im dying tonight
 had cold hands feet severe headache fatigue fever dislike bright lights nausea would think
 meningitis possible early symptoms symptoms appear appear order

Fig. 4. Example of a symptomatic manifestation found in tweets

Its purpose is to improve the annotation based on the ontology IDOMEN. It is built to be scalable, which will take into account the specificity of expression of social media. MeningVocab resource was built by collecting tweets related to meningitis. We have adopted the OWL-compliant Simple Knowledge Organization System (SKOS) formalism to develop this resource. This resource is built independently of IDOMEN which will make it easier to manage the evolution of the vocabulary. An alignment will be done with the labels of the ontology IDOMEN. After analyzing the collected tweets, we note (manually scrolls through the corpus of tweets) some terms are closed to the clinical symptoms (used as concepts in IDOMEN ontology).

In this version of the SKOS meningVocab vocabulary we have 10 *concepts*, 20 *prefLabel* (in English and French), and 60 *altLabel*. We made two collections in the SKOS resource: a disease-related collection called “*Symptoms*” and a context-bound collection called “*Contexts*”.

3.3 The Resource “*hybrid*” : IDOMEN Aligned with meningVocab

The alignment of the two resources and their alignment allows us to have a hybrid resource that we use. Indeed, we have 2 types of IDOMEN ontology concepts that are used. These are concepts related to clinical symptoms and concepts related to socio-climatic aspects. These concepts are aligned with the skos resource using the altLabel of these concepts that we match with the prefLabel of the meningVocab concepts (Table 2).

Table 2. Extract from The resource *“hybrid”* : IDOMEN aligned with meningVocab

IDOMEN		meningVocab			
Concept	altLabel	skos:Concept	skos:prefLabel@en	skos:prefLabel@fr	skos:altLabel@en
IDOMEN_0000040	fever	vocmen0101	fever	fièvre	high temperature; hot body, high body temp, body temp, acute fever
IDOMEN_0000031	headache	vocmen0102	headache	mal de tête	severe headache, migraine
IDOMEN_0000029	stiff neck	vocmen0103	stiff neck	raideur de la nuque	unable move neck, neck pain
IDOMEN_0000032	confusion	vocmen0104	confusion	confusion	lethargy, hallucination, irritability
IDOMEN_0000041	convulsion	vocmen0106	convulsion	convulsion	spasm, muscle pain, convulsing
IDOMEN_0000026	vomiting	vocmen0107	vomiting	vomissement	vomited, nausea, vomit
IDOMEN_0000028	photophobia	vocmen0108	photophobia	photophobie	dislike light, light sensitivity, eye sensitivity, sensitivity light

4 Strategy of Collecting and Filtering

In this section, we propose a strategy for collecting and filtering relevant messages from social networks as part of the health monitoring of meningitis epidemics based on termino-ontological resources. It is based on 4 steps: (1) collection and filtering, (2) pre-processing, (3) annotation of tweets and matching, and (4) enforcement of rules. The main challenge is to be able to extract relevant and meaningful terms related to meningitis from the huge amount of data produced on Twitter (Fig. 5).

4.1 Step 1: Collect and Filter Messages from Twitter

The collection: the daily tweets flow is estimated at just over 500 million tweets per day [9]. The Twitter Python API used to access tweets in real time. The streaming of tweets is then analyzed by retrieving those that are related to the vocabulary of IDOMEN ontology.

The filtering: we use the vocabularies of aligned resource to filter messages related to meningitis. The labels of ontology concepts are aligned with the terms of the SKOS resource. All the terms designating the concepts of the domain are collected in a list of term candidates using the Owlready API. Message filtering is carried out using two categories of filters that we apply sequentially:

1. Filter on clinical aspects;
2. Filter on the climatic aspects of meningitis.

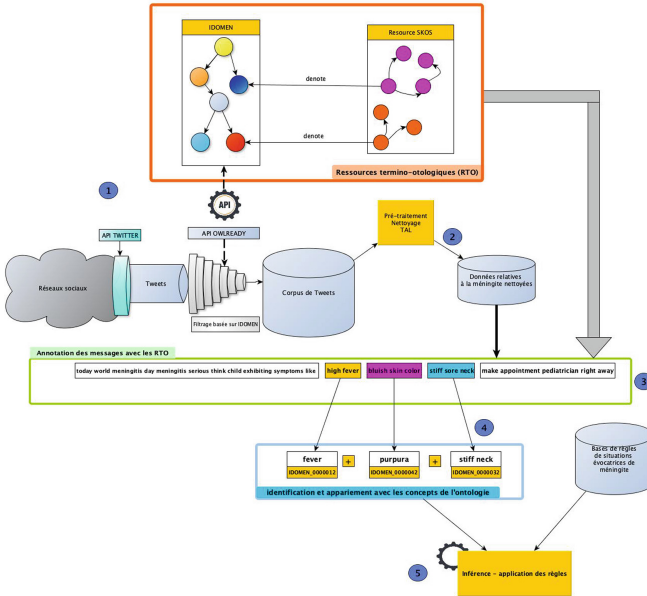


Fig. 5. Processing chain

Filter on Clinical Aspects

The filter based on the symptoms and signs of meningitis is a filter consisting of candidate terms related to the symptoms of meningitis. All subclasses of parent classes: “meningitis symptom” and “meningitis sign” in the ontology will serve as a filter on the clinical aspects. The subclasses are: “vomiting”, “burning of the fontanelle”, “photophobia”, “stiff neck”, “fever”, “headache” and “purpura”.

Filter on the Climatic Aspects Related to Meningitis

The evolution of meningitis (appearance, development, and intensity) is strongly linked to socio-demographic, economic, climatic and environmental factors [11]. It is therefore appropriate to consider these dimensions in the context of meningitis surveillance to detect events that may be risk factors for the spread of meningitis. We use filtering based on climatic and environmental events. The filter is based on the subclasses “climatic” and “environmental factors”. We therefore find the following terms: “humidity”, “wind speed”, “cold”, “rainfall”, “solar radiation”, “atmospheric pressure”, “temperature”, “dust”, “cloud dust”.

4.2 Step 2: Pre-treatment and Cleaning

NLP pre-processing techniques are then used to clean tweets from artifacts such as emoticons, RT (ReTweet) statements, URLs, unnecessary punctuation, etc. The decomposition of tweets into sentences is called tokenization. This operation allows us to split the tweet into sentences from which we delete the stopwords.

Some examples of stopwords in english (“by”, “than”, “could not”, “re”, “is not”, “on”, “my”, “would”, “up”, “we”, “doesn”, “is”, “doing”, “haven”, “an”, etc.) are words that do not add value to our analysis. We use the N-Gram technique, to break down tweets into N-gram. To do this, we use the Python TextBlob library, which allows us to process text data such as partial tagging, extraction of nominal phrases, sentiment analysis, classification, translation, etc. Each N-gram passes through the filters described above. If there is a match with one of the filter terms then it will be considered a relevant tweet.

4.3 Step 3: Annotation with RTOs

The annotation process consists of annotating the corpus of tweets based on the two resources: the IDOMEN ontology vocabulary and the controlled vocabulary of meningVocab. The purpose of this step is to identify a term or phrase and to be able to match it with an ontology concept using the SKOS resource “meningVocab”. We use existing relationships between terms to associate them with a symptom of meningitis.

4.4 Step 4: Application of the Domain Rules

As part of the medical diagnosis of meningitis, [12] have established rules to recognize situations suggestive of meningitis. These rules have been translated into SWRL using the IDOMEN ontology. After identifying or matching the symptoms contained in a tweet, the 5 rules below are applied to qualify and select the tweets to be processed (Fig. 6).

- Rule 1 : Fever + stiff neck + confusion \Rightarrow *suspectedCase*;
- Rule 2 : Fever + stiff neck + headache \Rightarrow *suspectedCase*;
- Rule 3 : Fever + purpura + headache \Rightarrow *suspectedCase*;
- Rule 4 : Fever + focal neurological signs \Rightarrow *suspectedCase*;
- Rule 5 : Fever + convulsions \Rightarrow *suspectedCase*;

```
autogen1:IDOMEN_0000023(?symA, ?y) ^ autogen1:IDOMEN_0000040(?symA) ^
autogen1:IDOMEN_0000023(?symB, ?y) ^ autogen1:IDOMEN_0000041(?symB) ->
autogen1:IDOMEN_0000021(?y)
```

Fig. 6. Rule 4 in IDOMEN : Fever + focal neurological signs = suspected case

5 Discussion and Conclusions

The Table 3 gives the results of the new annotations made on the same corpus of 1413 tweets with the SKOS meningVocab resource. At step 3 of our processing we will have 1780 annotations (with IDOMEN) and 324 additional terms

Table 3. Results of the annotation with the vocabulary of resources aligned

Terms of resources aligned	Number of occurrences
Hot body	0
High body temperature	0
Acute fever	2
Severe headache	58
Migraine	22
Unable move neck	1
Lethargy	10
Irritability	11
Hallucination	5
Focal neurologic	5
Spasm	3
Muscle pain	13
Dislike bright light	17
Light sensitivity	3
Eye sensitivity	4
Sensitivity light	17
Stain blood	13
Rash	52
Nausea	86
Total	324

(with meningVocab specific terms missing from IDOMEN), that is a total of 2104 annotations using the two termino-ontological resources . This result is not surprising since terms used by users are already in IDOMEN. The 324 annotations identify symptoms that are not usually designated by scientific terms such as “*photophobia*” appears very rarely while terms such as “*dislike bright light*”, “*light sensitivity*”, “*eye sensitivity*” or “*sensitivity light*” are used to describe it. Using the two resources, we obtained 2104 annotations.

In this article we have presented a strategy for collecting and filtering data related to meningitis using two termino-ontological resources: IDOMEN and meningVocab. Our approach takes advantage of the complementarity of these two resources. The use of the SKOS resource allows to enrich the one obtained from IDOMEN. We plan to evaluate the relevance of tweets selected. In a first approach, the choice of terms used to build the meningVocab vocabulary was done manually by going through the initial corpus. In future work we will use automatic learning techniques (including a lexical extension model on the corpus) to propose better term candidates for the SKOS resource .

References

1. Béré, W.R.C., Camara, G., Malo, S., Lo, M., Ouaro, S.: IDOMEN: an extension of infectious disease ontology for MENingitis. *Stud. Health Technol. Inform.* **264**, 313–317 (2019). <https://doi.org/10.3233/SHTI190234>. <http://europepmc.org/abstract/med/31437936>
2. Broniatowski, D.A., Paul, M.J., Dredze, M.: National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. *PLoS ONE* **8**(12), e83672 (2013). <https://doi.org/10.1371/journal.pone.0083672>. <https://dx.plos.org/10.1371/journal.pone.0083672>
3. Camara, G., Després, S., Djedidi, R., Lo, M.: Vers une ontologie des processus de propagation des maladies infectieuses. In: 23èmes journées francophones d'ingénierie des connaissances, pp. 99–114. IC 2012, June 2012
4. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics, pp. 115–122. SOMA 2010, Association for Computing Machinery, Washington D.C., District of Columbia, July 2010. <https://doi.org/10.1145/1964858.1964874>
5. Cunningham, H., Tablan, V., Roberts, A., Bontcheva, K.: Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLOS Comput. Biol.* **9**(2), e1002854 (2013). <https://doi.org/10.1371/journal.pcbi.1002854>. <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002854>
6. Dredze, M., Paul, M., Bergsma, S., Tran, H.: Carmen: a twitter geolocation system with applications to public health. In: AAAI Workshop - Technical Report, pp. 20–24, January 2013
7. Eysenbach, G.: Infodemiology and infoveillance framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J. Med. Internet Res.* **11**(1), e11 (2009)
8. Gomez-Perez, A., Fernández-López, M., Corcho, O.: *Ontological Engineering: with examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, London (2006). <https://doi.org/10.1007/b97353>
9. Liu, S., Young, S.D.: A survey of social media data analysis for physical activity surveillance. *J. Forensic Legal Med.* **57**, 33–36 (2018). <https://doi.org/10.1016/j.jflm.2016.10.019>. <https://linkinghub.elsevier.com/retrieve/pii/S1752928X1630141X>
10. Vilain, P., Menudier, L., Filleul, L.: Twitter: a complementary tool to monitor seasonal influenza epidemic in France? *Online J. Public Health Inform.* **11**(1), e296 (2019). <https://doi.org/10.5210/ojphi.v11i1.9724>. <https://journals.uic.edu/ojs/index.php/ojphi/article/view/9724>
11. Yaka, P., Sultan, B., Broutin, H., Janicot, S., Philippon, S., Fourquet, N.: Relationships between climate and year-to-year variability in meningitis outbreaks: a case study in Burkina Faso and Niger. *Int. J. Health Geograph.* **7**(1), 34 (2008). <https://doi.org/10.1186/1476-072X-7-34>. <http://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-7-34>
12. Zender, H.O., Olivier, P., Genné, D.: Méningites bactériennes communautaires aiguës chez l'adulte (2009). <https://www.revmed.ch/RMS/2009/RMS-220/Meningites-bacteriennes-communautaires-aigues-chez-l-adulte>