Enjeux et opportunités de la fouille textes pour stimuler la recherche pluridisciplinaire

Mathieu ROCHE

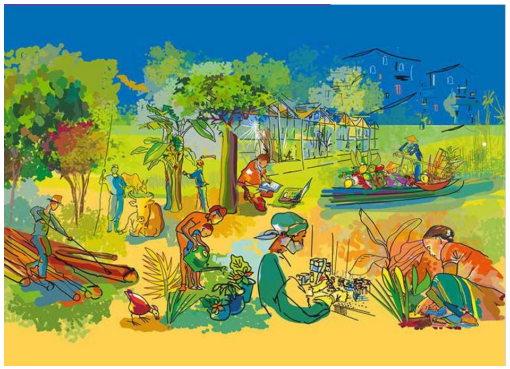
mathieu.roche@cirad.fr

Cirad, UMR TETIS, Montpellier

Contexte pluridisciplinaire

UMR TETIS

Cirad



#DigitAg

asds
THE AFRICAN SOCIETY IN DIGITAL SCIENCES

Contexte pluridisciplinaire

On ne peut faire concurrence aux GAFAM en termes de **données** et **modèles**

Les travaux **pluridisciplinaires** peuvent constituer une opportunité pour la **recherche académique**

Géographes

Linguistes

Territoire

Agriculture

One Health

Sociologues

Agronomes

Vétérinaires

Plan de la présentation

1. Comment les **projets pluridisciplinaires** permettent de construire des **démarches génériques** ?

2.

Données

Modèles

Thématique

Fouille de textes

3.

4. Quels sont les **nouveaux défis pluridisciplinaires** en particulier dans les pays du Sud ?

Plan de la présentation

1. Comment les **projets pluridisciplinaires** permettent de construire des **démarches génériques** ?

→ **Données :** Annotation, diffusion et valorisation de corpus

→ **Méthode :** Pipeline et démarche générique

Plan de la présentation

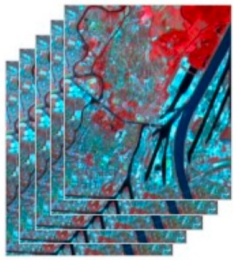
1. Comment les **projets pluridisciplinaires** permettent de construire des **démarches génériques** ?

→ **Données :** Annotation, diffusion et valorisation de corpus

→ **Méthode :** Pipeline et démarche générique

Corpus de segments

Territoire

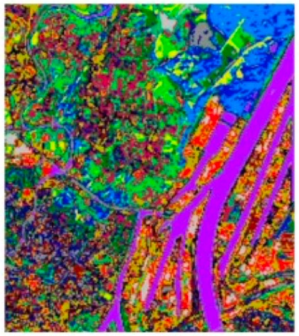


images satellites

Sémantisation des données satellitaires

contraintes

documents pertinents



clusters labellisés

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10

Extraction de règles à partir de documents d'urbanisme liés aux sites de recherche qui nous intéressent

Corpus de segments

[Koptelov *et al.*, Scientific Data 2023]

Territoire

scientific **data**

OPEN

DATA DESCRIPTOR

Check for updates

A manually annotated corpus in French for the study of urbanization and the natural risk prevention

Maksim Koptelov[1,2,3] ✉, Margaux Holveck[4], Bruno Cremilleux[1], Justine Reynaud[1], Mathieu Roche[3,5] ✉ & Maguelonne Teisseire[2,3]

Land artificialization is a serious problem of civilization. Urban planning and natural risk management are aimed to improve it. In France, these practices operate the Local Land Plans (PLU – Plan Local d'Urbanisme) and the Natural risk prevention plans (PPRn – Plan de Prévention des Risques naturels) containing land use rules. To facilitate automatic extraction of the rules, we manually annotated a number of those documents concerning Montpellier, a rapidly evolving agglomeration exposed to natural risks. We defined a format for labeled examples in which each entry includes title and subtitle. In addition, we proposed a hierarchical representation of class labels to generalize the use of our corpus. Our corpus, consisting of 1934 textual segments, each of which labeled by one of the 4 classes (Verifiable, Non-verifiable, Informative and Not pertinent) is the first corpus in the French language in the fields of urban planning and natural risk management. Along with presenting the corpus, we tested a state-of-the-art approach for text classification to demonstrate its usability for automatic rule extraction.

Hérelles
Un cadre collaboratif unifié pour l'analyse interactive de données temporelles

Hérelles ANR Project

(Université de Strasbourg, INRAE, Université d'Orléans, Université de Caen, AgroParisTech)

Recherche Data Gouv > Université de Strasbourg > iCUBE - Science des Données et Connaissances - UMR 7357 > Hérelles ANR Project >

Segments textuels consolidés - Consolidated Textual Segments - Hérelles Project

Version 5.0

Holveck, Margaux; Koptelov, Maksim; Roche, Mathieu; Teisseire, Maguelonne, 2023, "Segments textuels consolidés - Consolidated Textual Segments - Hérelles Project", https://doi.org/10.57745/XIVJ65, Recherche Data Gouv, V5, UNF:6:9rR4WHi1boyTJSzj33qDfg== [fileUNF]

Citer le jeu de données

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Modalités d'accès au jeu de données

Contact | Partager

Statistiques d'utilisation sur les jeux de données

901 consultations

490 téléchargements

0 citation

Description

(English version below) L'un des objectifs du projet Hérelles est de trouver de nouveaux mécanismes afin de faciliter l'étiquetage (ou sémantisation) des clusters issus des séries temporelles d'images satellite. Pour y parvenir, une solution proposée est d'associer des éléments textuels d'intérêt (adéquation avec la thématique d'étude, et le périmètre spatio-temporel des séries temporelles) aux données satellite. Ce jeu de donnée est une version consolidée du jeux de donnée "Segments Textuels Hérelles". Il présente un corpus thématique préalablement récolté et annoté manuellement ainsi que le code et les résultats d'une méthode d'extraction automatique des éléments textuels d'intérêt. Il comprend les éléments suivants :

I **Article 1 : Occupations ou utilisations du sol interdites**

1 **1) Dans tous les secteurs :**

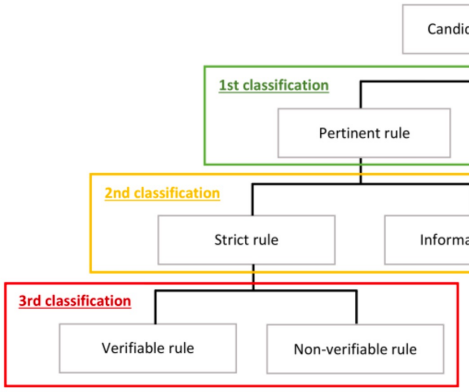
- i Les constructions destinées à l'industrie, à l'artisanat et à la fonction d'entrepôt.
- ii Les pylônes et poteaux, supports d'enseignes et d'antennes d'émission ou de réception de signaux radioélectriques.
- iii Les installations classées pour la protection de l'environnement soumises à déclaration ou à autorisation, autres que celles visées à l'article 2 paragraphe 2).
- iv Les constructions destinées à l'habitat, au commerce, au bureau, à l'hébergement hôtelier autres que celles visées à l'article 2, paragraphe 2).
- v Les constructions ou installations d'intérêt collectif autres que celles visées à l'article 2, paragraphe 2) et 3.

2 **2) Dans les périmètres en bordure des cours d'eau définis dans les annexes sanitaires du PLU :**

- i Les occupations et utilisations autres que celles visées à l'article 2, paragraphe 3).

3 **3) Dans les périmètres des secteurs particuliers de risque d'inondation délimités dans les documents graphiques du règlement :**

- i Les occupations et utilisations autres que celles visées à l'article 2, paragraphe 4).



CamemBERT **F1 score 91%**

ML approch (TF-IDF, SGD) **Accuracy pondérée 93%**

CamemBERT + augmentation des données **Accuracy pondérée 93%**

L'augmentation améliore les résultats de 6 % (F1 score classe Vérifiable)

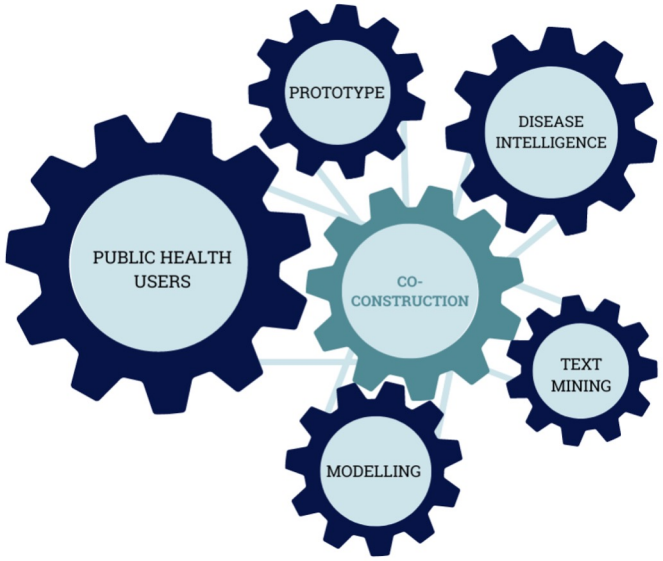
MOnitoring Outbreaks for Disease surveillance in a data science context

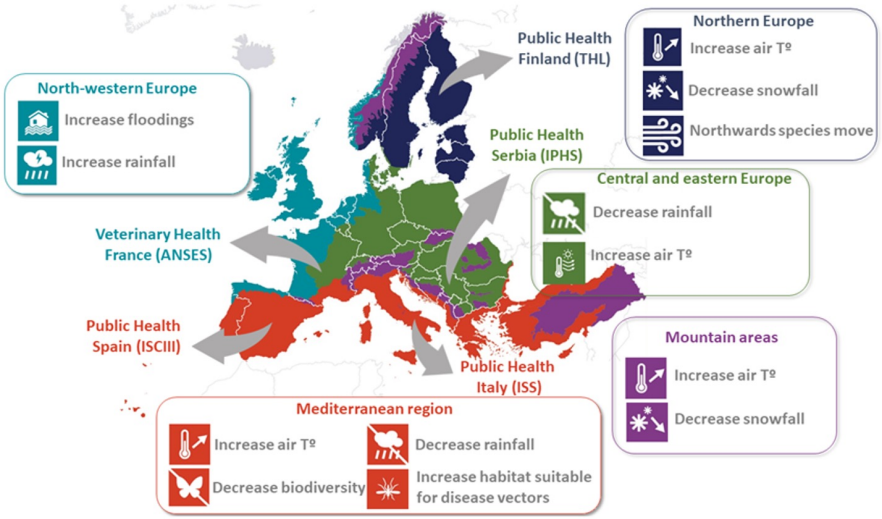
One Health

Horizon 2020 (2020-2024)

Partners: 14 mil, 25 partners, 13 countries

Targets: Health (PH/AH/OH) agencies in Europe and epidemic intelligence practitioners





Corpus de phrases

[Valentin *et al.*, Scientific Data 2022]

One Health

scientific **data**

OPEN

DATA DESCRIPTOR

Elaboration of a new framework for fine-grained epidemiological annotation

Sarah Valentin[1,2,3,4], Elena Arsevska[2,5], Aline Vilain[6], Valérie De Waele[7], Renaud Lancelot[2,5] & Mathieu Roche[1,5]

Event-based surveillance (EBS) gathers information from a variety of data sources, including online news articles. Unlike the data from formal reporting, the EBS data are not structured, and their interpretation can overwhelm epidemic intelligence (EI) capacities in terms of available human resources. Therefore, diverse EBS systems that automatically process (all or part of) the acquired nonstructured data from online news articles have been developed. These EBS systems (e.g., GPHIN, HealthMap, MedISys, ProMED, PADI-web) can use annotated data to improve the surveillance systems. This paper describes a framework for the annotation of epidemiological information in animal disease-related news articles. We provide annotation guidelines that are generic and applicable to both animal and zoonotic infectious diseases, regardless of the pathogen involved or its mode of transmission (e.g., vector-borne, airborne, by contact). The framework relies on the successive annotation of all the sentences from a news article. The annotator evaluates the sentences in a specific epidemiological context, corresponding to the publication date of the news article.

Dataverse — Search — User Guide — Support — Log In

tetis — UMR TETIS — Territoires, Environnement, Télédétection et Information Spatiale

(Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE)

CIRAD Dataverse > UMR TETIS >

Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus

Version 3.0

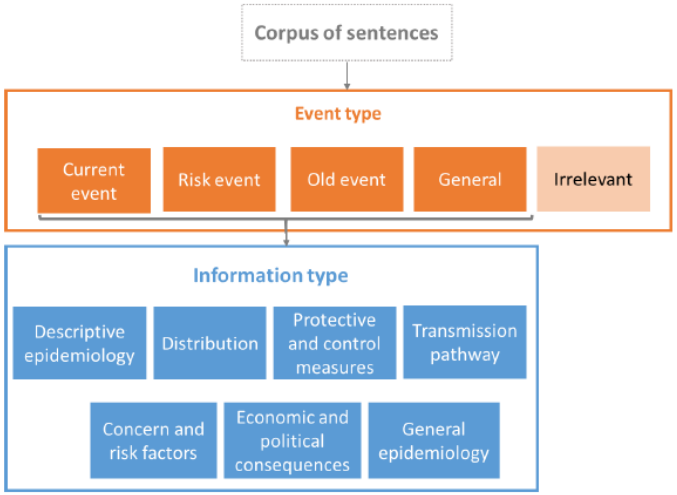
Valentin Sarah; De Waele Valérie; Vilain Aline; Arsevska Elena; Lancelot Renaud; Roche Mathieu, 2019, "Annotation of epidemiological information in animal disease-related news articles: guidelines and manually labelled corpus", https://doi.org/10.18167/DVN1/YGAKNB, CIRAD Dataverse, V3, UNF: 6:H+qzG30RSQ4fWYYA2UBwEQ== [fileUNF]

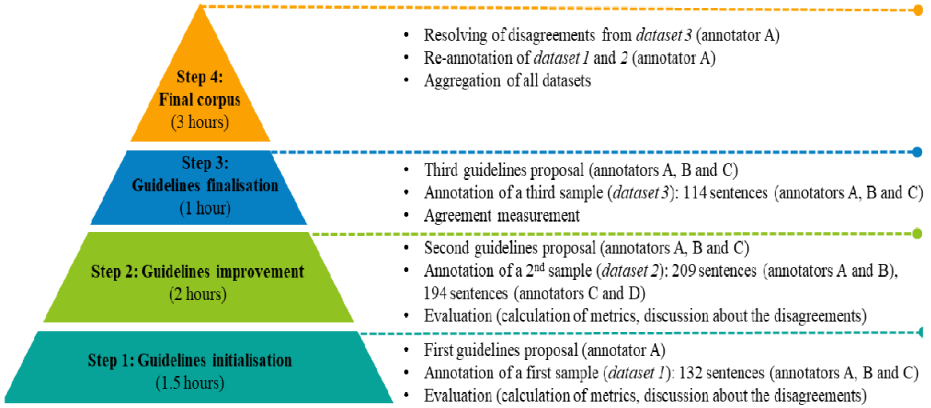
Cite Dataset — Learn about Data Citation Standards.

Access Dataset — Contact Owner — Share

Dataset Metrics

111 Downloads





Corpus de dépêches

[Arinik *et al.*, Data in Brief 2023]

One Health

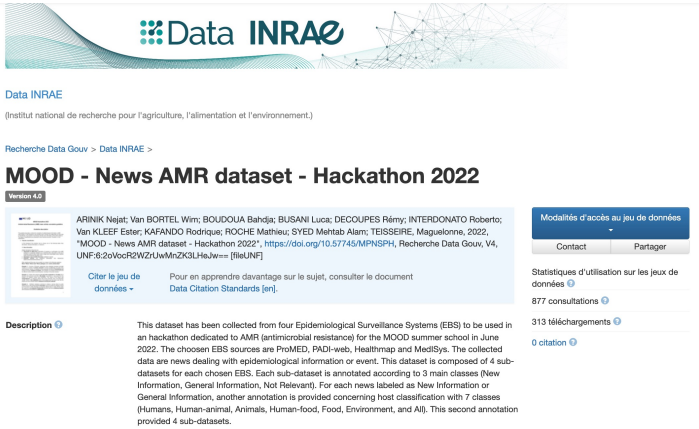
Data in Brief

Volume 46, February 2023, 108870

Data Article

An annotated dataset for event-based surveillance of antimicrobial resistance

Nejat Arınık [a, c], Wim Van Bortel [d], Bahdja Boudoua [a, c], Luca Busani [e], Rémy Decoupes [a, c], Roberto Interdonato [b, c], Rodrique Kafando [a, c], Esther van Kleef [f], Mathieu Roche [b, c], Mehtab Alam Syed [b, c], Maguelonne Teisseire [a, c]



Data INRAE

Data INRAE

(Institut national de recherche pour l'agriculture, l'alimentation et l'environnement.)

Recherche Data Gouv > Data INRAE >

MOOD - News AMR dataset - Hackathon 2022

Version 4.0

ARINIK Nejat; Van BORTEL Wim; BOUDOUA Bahdja; BUSANI Luca; DECOUPES Rémy; INTERDONATO Roberto; Van KLEEF Ester; KAFANDO Rodrique; ROCHE Mathieu; SYED Mehtab Alam; TEISSEIRE, Maguelonne, 2022, "MOOD - News AMR dataset - Hackathon 2022", https://doi.org/10.57745/MPNSPH, Recherche Data Gouv, V4, UNF:6:2oVocR2WZrUwMnZK3LHeJw== [fileUNF]

Citer le jeu de données

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Modalités d'accès au jeu de données

Contact | Partager

Statistiques d'utilisation sur les jeux de données

877 consultations

313 téléchargements

0 citation

Description

This dataset has been collected from four Epidemiological Surveillance Systems (EBS) to be used in an hackathon dedicated to AMR (antimicrobial resistance) for the MOOD summer school in June 2022. The choosen EBS sources are ProMED, PADI-web, Healthmap and MedISys. The collected data are news dealing with epidemiological information or event. This dataset is composed of 4 sub-datasets for each chosen EBS. Each sub-dataset is annotated according to 3 main classes (New Information, General Information, Not Relevant). For each news labeled as New Information or General Information, another annotation is provided concerning host classification with 7 classes (Humans, Human-animal, Animals, Human-food, Food, Environment, and All). This second annotation provided 4 sub-datasets.

MEDISYS

ProMED mail

Program for Monitoring Emerging Diseases

HealthMap

Global Health, Local Information

twitter

PADI WEB

Hackathon corpus

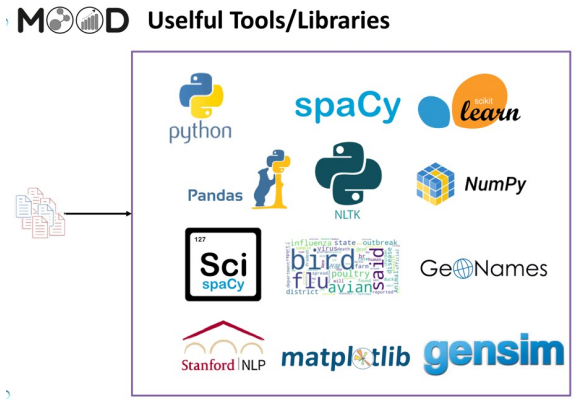
(AMR)

Corpus de dépêches

One Health

First step: Definition of classes (Guideline) based on a sample of 25 news with 2-rounds of annotation. Annotation done by 3 epidemiologists (Esther van Kleef, Wim Van Bortel, Luca Busani) and supervised by a text-miner (Mathieu Roche)

Second Step: Annotation of datasets by computer scientists (Nejat Arinik, Mehtab Alam Syed, Maguelonne Teisseire, Remy Decoupes, Roberto interdonato) and supervised by an epidemiologist (Bahdja El Boudoua)



Plan de la présentation

1. Comment les **projets pluridisciplinaires** permettent de construire des **démarches génériques** ?

→ **Données :** Annotation, diffusion et valorisation de corpus

→ **Méthode :** Pipeline et démarche générique

PADI-web

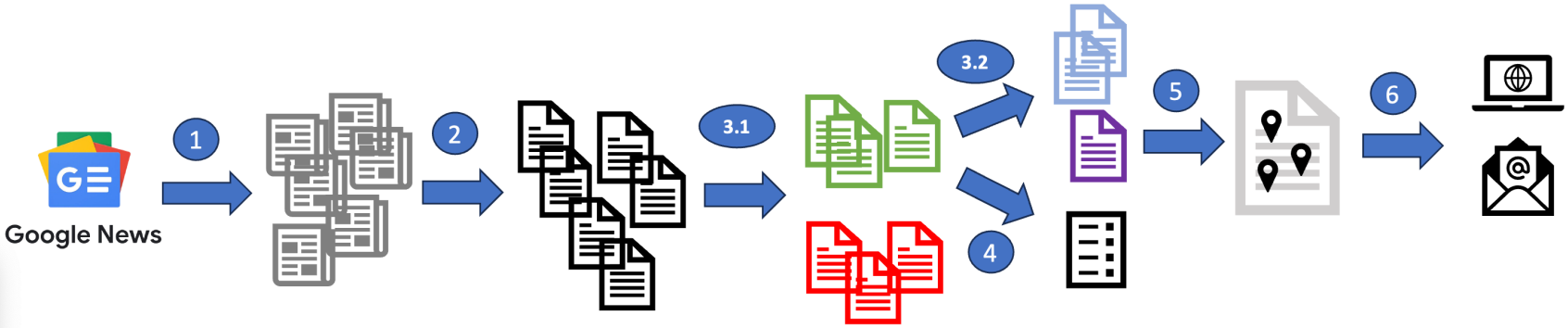
[Valentin *et al.*, One Health 2021]

One Health

PADI-web *Animal Health*

PADI-web *Plant Health*

PADI-web *Public Health*



Step 1: Data collection
- Multlingual articles

Step 2: Processing
- Text cleaning
- Language detection
- Translation into English

Step 3: Document classification
- **Step 3.1:** Relevant/irrelevant documents
- **Step 3.2:** Topic of documents

Step 4: Sentence classification
- Topic of sentences

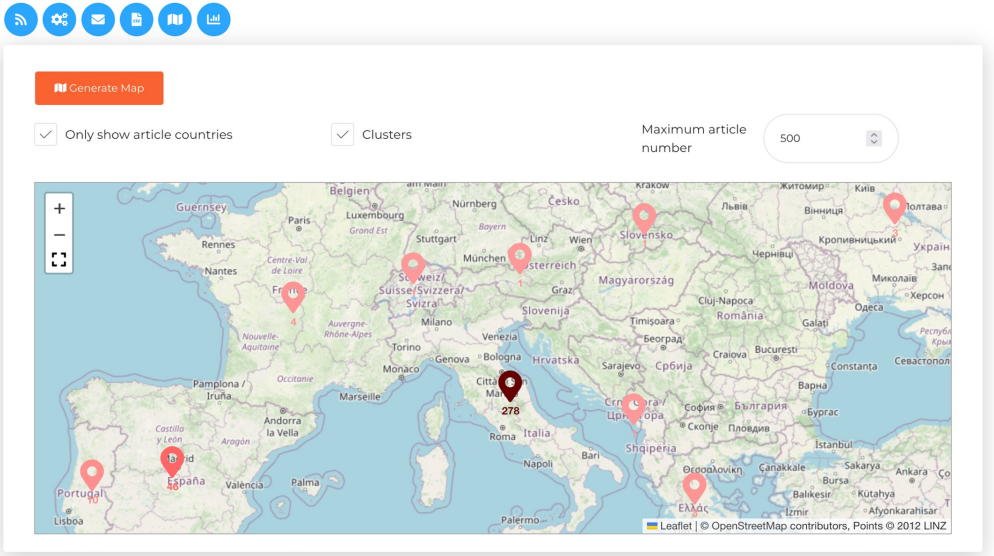
Step 5: Information extraction
- Disease
- Location
- Host
- Symptom
- etc.

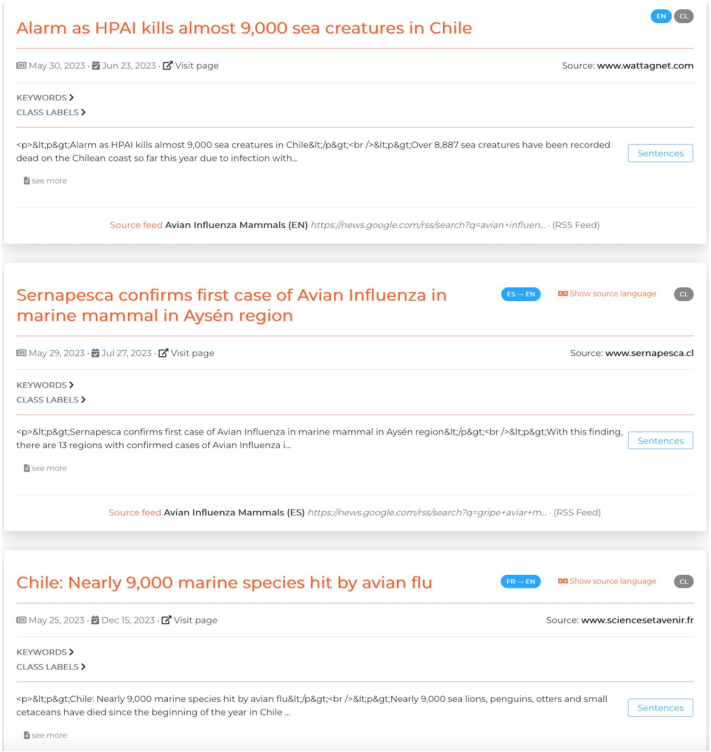
Step 6: Output for end-users
- Visualisations
- Notifications

PADI-web

[Valentin *et al.*, One Health 2021]

One Health



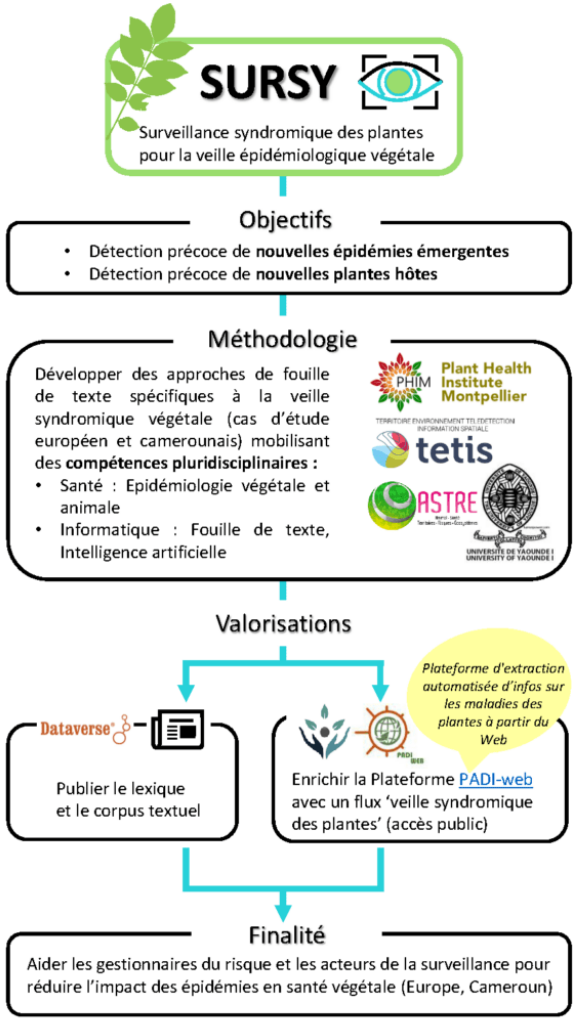


PADI-web

Santé végétale

[Roche *et al.*, CAISE 2024]

One Health



Random Forest:
Accuracy at 0.79 (5-fold cross-validation process)

RoBERTa:

Class	Accuracy	Precision	Recall
Communication/popularization	0.85	0.84	0.85
Genetic and molecular scale	0.90	0.89	0.90
Health statuses and interceptions	0.87	0.86	0.87
Fighting measures	0.85	0.84	0.85
Methods of analysis and detection	0.97	0.96	0.97
Regulation	0.94	0.88	0.94
Epidemiological, socio-economic and environmental risk	0.96	0.95	0.96
Monitoring	0.89	0.80	0.89

Plan de la présentation

2. Comment les **thématiques** peuvent **engendrer de nouveaux verrous scientifiques** pour la fouille de textes ?

→ **Données :** Données de spécialité

→ **Méthode :** Production de nouveaux modèles

2.

Thématique  **Fouille de textes**

Mise en lien de variables

Agriculture

Contents lists available at ScienceDirect

Data in Brief

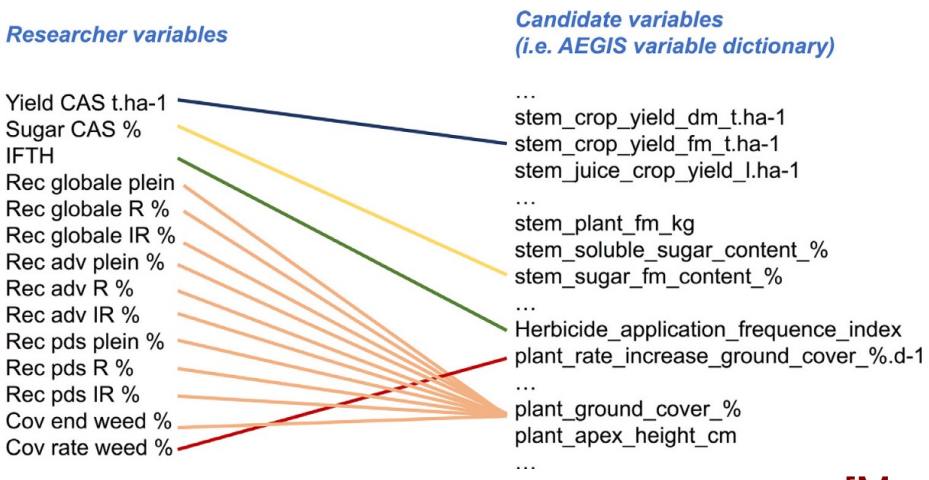
journal homepage: www.elsevier.com/locate/dib

Data Article

Experimental variables in sugarcane intercropping in Reunion Island for data matching

Sandrine Auzoux[a,c], Billy Ngaba[a,c], Mathias Christina[a,c], Benjamin Heuclin[a,c], Mathieu Roche[b,c,*]

[a] UR AIDA (Agroecology and sustainable intensification of annual crops), University of Montpellier, CIRAD, La Réunion, France
[b] UMR TETIS (Land, Environment, Remote Sensing and Spatial Information), University of Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[c] French Agricultural Research for Development (CIRAD), France



[Mecchour *et al.*, EGC 2024]

Plan de la présentation

2. Comment les **thématiques** peuvent **engendrer de nouveaux verrous scientifiques** pour la fouille de textes ?

→ **Données :** Données de spécialité

→ **Méthode :** Production de nouveaux modèles

Prompt engineering

Zero-shot learning

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

One-shot learning

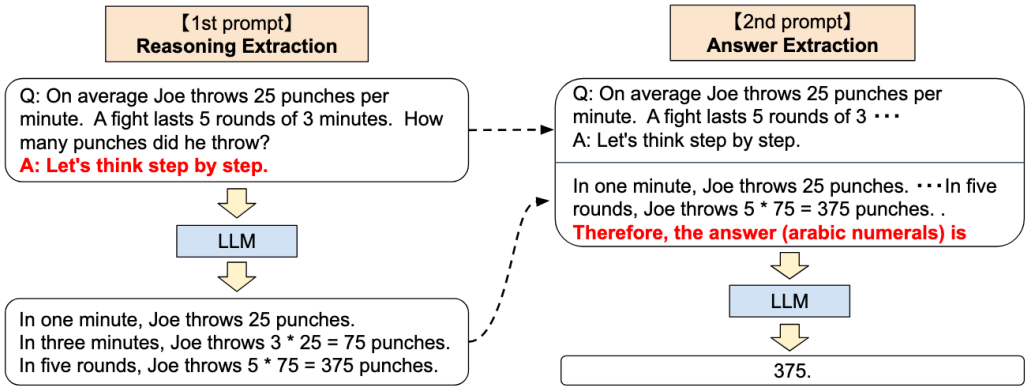
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot learning

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

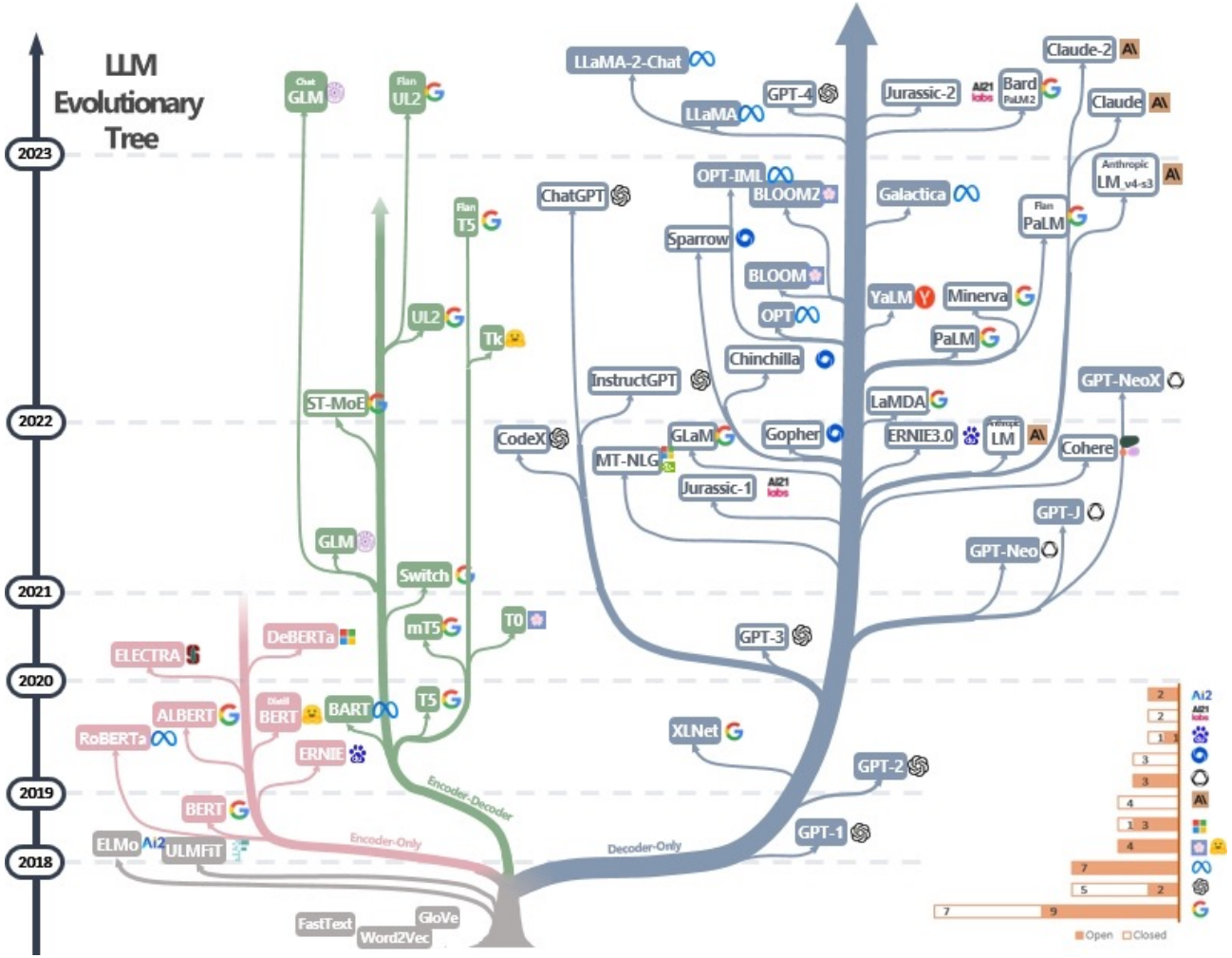
CoT (chain of thought prompting / chaine de pensée) : *Feed LLMs with the step-by-step reasoning*

Zero-shot-CoT : *Let's think step by step before each answer*



[Kojima *et al.*, NIPS 2022]

Des modèles foisonnants

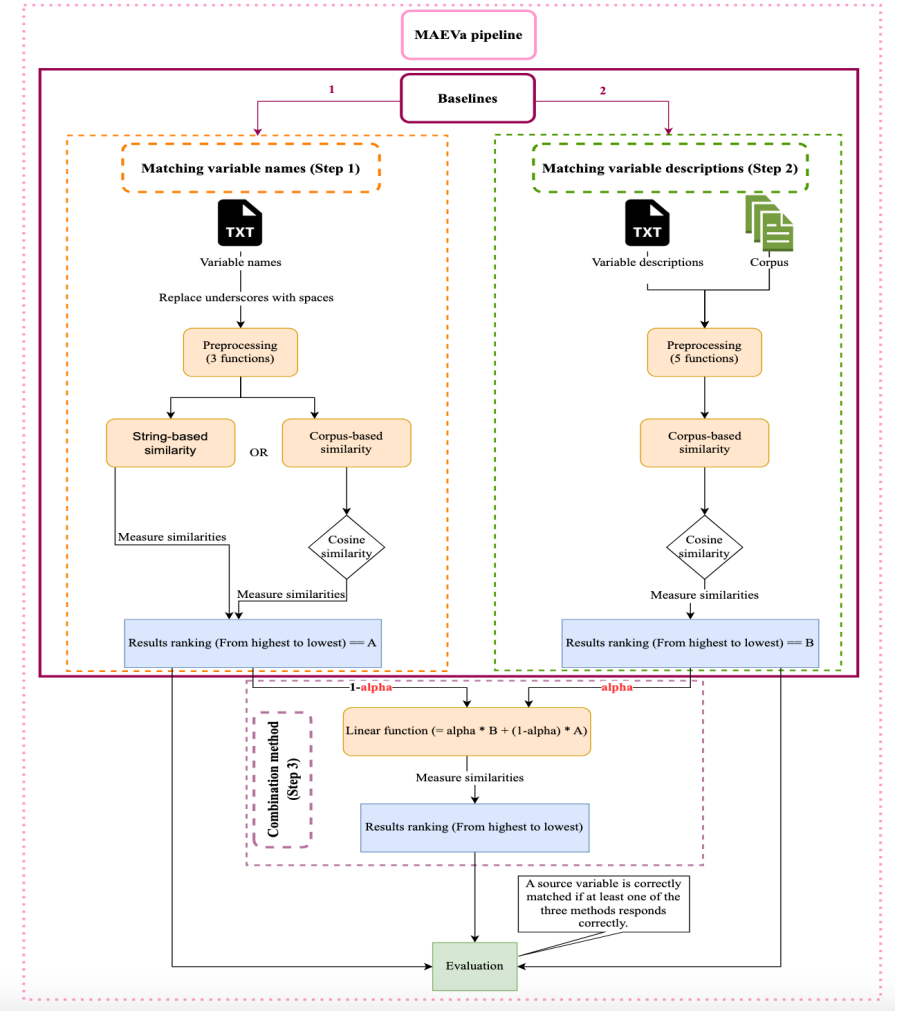


Les **grands modèles de langues** (LLM) sont fondés sur des milliards de paramètres, là où BERT en possède 340 millions

Source: https://github.com/ Mooler0410/LLMsPracticalGuide

Agroécologie

Agriculture



Method	CMP@1	CMP@3	CMP@5	CMP@10
Jaro-Winkler Similarity	8.33%	17.86%	22.62%	30.95%
Overlap Coefficient	8.33%	21.43%	32.14%	42.86%
Sorensen-Dice Index	2.38%	3.57%	3.57%	5.95%
Tversky Index	26.19%	32.14%	45.24%	60.71%
BERT-base (2 HLs) + cosine similarity	11.90%	28.57%	36.90%	53.57%
BERT-base (2 HLs) + Multi-Head attention (256 Hs, DT=0.1 and UDW) + cosine similarity	**30.90%**	**40.33%**	**51.00%**	**64.69%**

Plan de la présentation

3. Comment les **travaux de fouille de textes** peuvent alimenter les **travaux pluridisciplinaires** ?

→ **Données :** données qui questionnent…

→ **Méthode :** Production de nouvelles approches

Thématique  **Fouille de textes**

3.

Plan de la présentation

3. Comment les **travaux de fouille de textes** peuvent alimenter les **travaux pluridisciplinaires** ?

→ **Données :** données qui questionnent…

→ **Méthode :** Production de nouvelles approches

Données qui questionnent…

Agriculture

Comment **mieux comprendre** la sécurité alimentaire et les sentiments

One Health

Comment **mieux interpréter** les indicateurs épidémiologiques

Données qui questionnent…

Quelques défis :

Hétérogénéité des données
→ Langues vernaculaires *vs.* textes scientifiques [Barreaux & Besagni, LREC 2020]
→ etc.

Traitement d'objets des publications scientifiques :
→ Théorèmes [Mishra *et al.*, JCDL 2024]
→ Tableaux [Lentschat *et al.*, ESWA 2022]
→ etc.

Lien entre données textuelles et autres types de données :
→ Représentation unifiée textes et images satellitaires [Neptune & Mothe, CBMI 2023]
→ etc.

Plan de la présentation

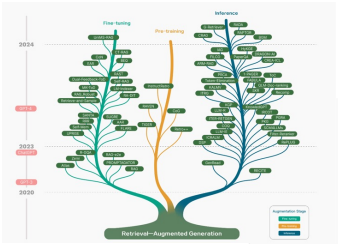
3. Comment les **travaux de fouille de textes** peuvent alimenter les **travaux pluridisciplinaires** ?

→ **Données :** Données originales et complexes

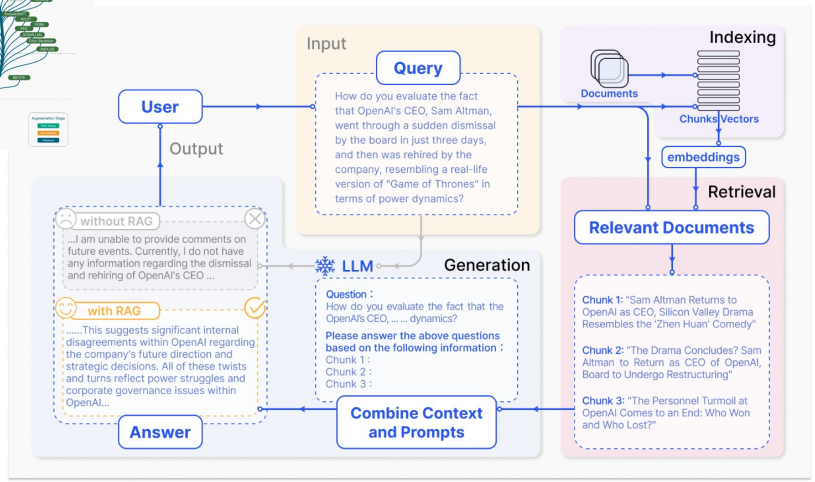
→ **Méthode :** Production de nouvelles approches

Fine-tuning vs. approche RAG

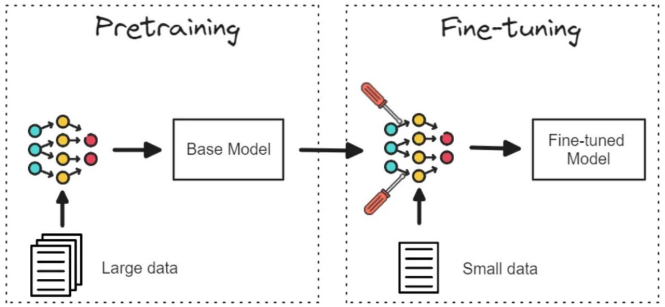
(Retrieval Augmented Generation)



RAG



Fine-Tuning



Source: https://medium.com/@prasadmahamulkar

[Gao *et al.*, arxiv 2024]

RAG vs. Fine-Tuning

Agriculture

Model	RAG	Fine-tuning
Cost – input token size	Increased Prompt Size	Minimal
Cost – output token size	More verbose, harder to steer	Precise, tuned for brevity
Initial cost	Low – creating embeddings	High – fine-tuning
Accuracy	Effective	Effective
New Knowledge	If data is in context	New skill in domain

Model	Fine-tuned	Accuracy	+RAG
Llama-2-chat 13B		76% ±2%	75% ±2%
Vicuna		72% ±2%	79% ±2%
GPT-4		75% ±3%	80% ±4%
Llama2 13B	✓	68% ±3%	77% ±2%
GPT-4	✓	81%±5%	86% ±2%

[Balaguer *et al.*, arxiv 2024]

Que disent les descripteurs textuels ?

Agriculture

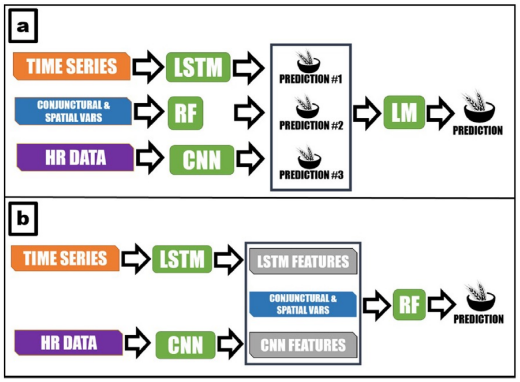
Utiliser les descripteurs textuels pour expliquer :
sémantisation spatiale

Data

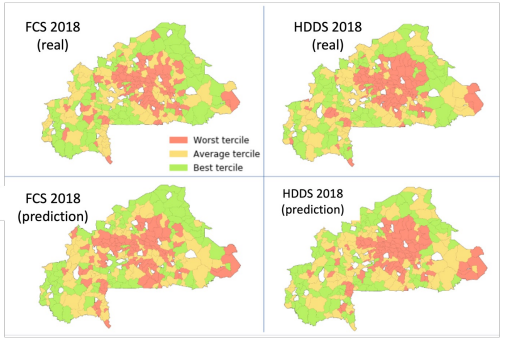
Summary of the datasets.

Variable	Resolution	Frequency	Source	Scaling up
Time series [several values per year; one value per commune] [70 vars]				
Smoothed brightness temperature (SMT) [14 vars]	4 km	7 days	National Oceanic and Atmospheric Administration (NOAA)	Maximum
Rainfall [14 vars]	6 km	10 days	Tropical Rainfall Measuring Mission (TRMM)	Sum
Average minimum and maximum temperatures [2 × 14 vars]	21 km	1 month	WorldClim	Mean
Maize price [14 vars]	64 markets	1 month	Société Nationale de Gestion du Stock de Sécurité alimentaire (SONAGESS)	K-nearest neighbour interpolation
Conjunctural data [one value per year; one value per commune] [20 vars]				
Meteorological data [7 vars]	10 stations	1 year	Knoema platform	K-nearest neighbour interpolation
Population density [4 vars]	100 m	1 year	Afripop	Spatial autocorrelation 2 km and 5 km, Gini, entropy
Economic data [7 vars]	Country	1 year	World Bank	Country value
Normalised difference vegetation index [2 vars]	250 m	1 year	Modis	Mean
Spatial data [one value per commune] [13 vars]				
Hospitals, schools [2 vars]	Point vectors	2018	Open Street Map	Count
Violent events [4 vars]	Point vectors	2018	Armed Conflict Location & Event Data Project (ACLED)	Count
Soil quality [3 vars]	1 km	2008	Food and Agriculture Organization (FAO)	Mean
Waterways [2 vars]	Line vectors	2008	Digital Chart of the World (DCW)	Count, length
Elevation data [2 vars]	1 km	2018	Consultative Group on International Agricultural Research (CGIAR)	Maximum, variance
High spatial resolution data [several values per commune] [4 vars]				
Population density	100 m	1 year	Afripop	CNN
Land cover (crops, forests, building areas)	20 m	2016	European Space Agency	CNN

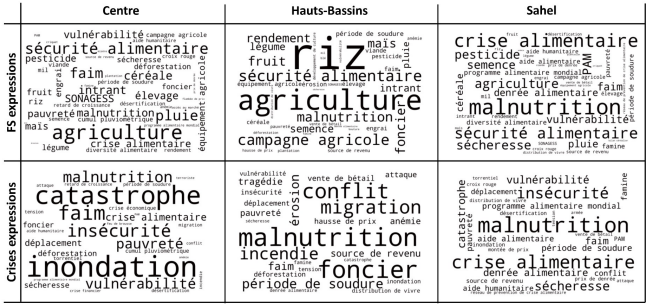
Process

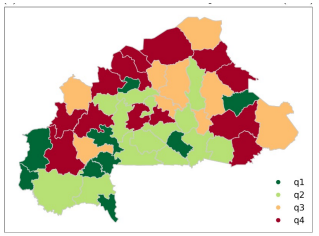


Output



[Deleglise *et al.* ESWA'2022]





Que disent les descripteurs textuels ?

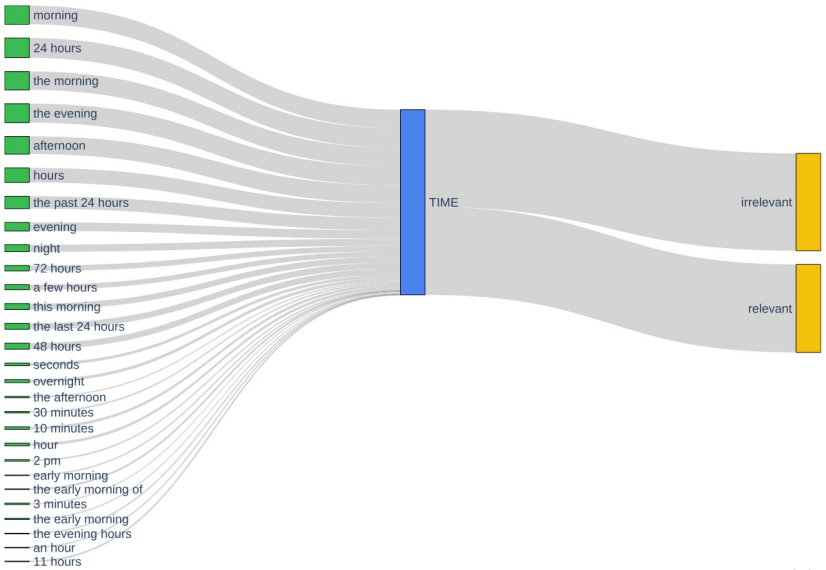
One Health

Utiliser les descripteurs textuels pour expliquer

Model	PADI-Web$_{long}$ Prec%	PADI-Web$_{long}$ Rec%	PADI-Web$_{long}$ F_1%	PADI-Web$_{XL}$ Prec%	PADI-Web$_{XL}$ Rec%	PADI-Web$_{XL}$ F_1%
PubMedBERT	91.36	92.72	92.03	**92.33**	93.52	92.92
BioELECTRA	56.42	**99.95**	72.14	60.33	96.57	74.26
BERT	56.36	98.27	71.63	56.64	**99.91**	72.32
SciBERT	90.35	91.65	90.99	91.46	91.82	91.63
EpidBioBERT	90.95	92.95	91.93	**92.33**	93.28	92.8
EpidBioELECTRA	**92.15**	93.49	**92.81**	**92.33**	94.62	**93.46**
ClinicalBERT	89.09	91.65	90.99	89.96	90.82	90.38

[Menya *et al.* ESWA 2024]

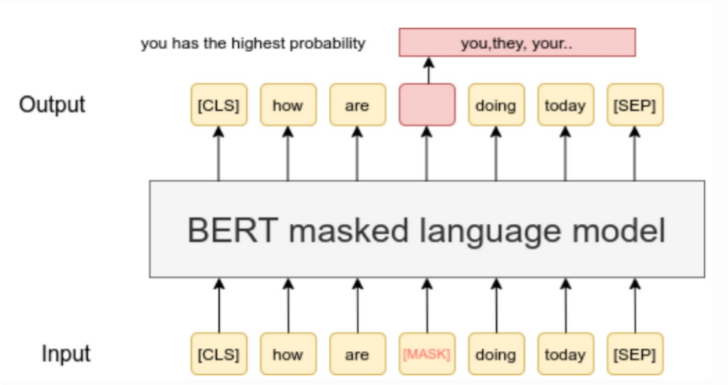
...the sardinia region is trying to get out of the nightmare represented by african swine fever, just as the rest of italy is on alert. in all on the island there are only three outbreaks in breeding and since 2019 there have been no more cases in pigs illegally kept in the wild and while among wild animals the virus has no longer been found, the island tries to reappear on the pork market beyond regional borders. the recent action of councillor murgia the regional councillor of agriculture, gabriella murgia, with the other colleagues of the agricultural policies commission of the state regions conference, met on 10 march...



Que disent les descripteurs textuels ?

Utiliser les descripteurs textuels pour guider :

Masked Language Model



Sélection des tokens peut être ciblée par :

- Des critères statistiques (confiance des algorithmes, TF-IDF)
- Les entités nommées extraites

[Gu *et al.* EMNLP 2020 ; Pergola *et al.* EACL 2021 ; Belfathi *et al.* arxiv 2024]

Que disent les descripteurs textuels ?

Utiliser les descripteurs textuels pour guider :

Masquage ciblé pour les modèles et des tâches de reconnaissance d'entités nommées

One Health

Could KeyWord Masking Strategy Improve Language Model?

Mariya Borovikova[1,3(✉)], Arnaud Ferré[1], Robert Bossy[1], Mathieu Roche[2,3], and Claire Nédellec[1]

[1] MaIAGE, Université Paris-Saclay, INRAE, Domaine de Vilvert, 78352 Jouy-en-Josas, France
mariya.borovikova@universite-paris-saclay.fr
[2] CIRAD, 34398 Montpellier, France
[3] TETIS, Univ. Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, 34090 Montpellier, France

[Borovikova *et al.* NLDB 2023]

Masquage ciblé pour l'augmentation de données

Territoire

SN Computer Science (2025) 6:115
https://doi.org/10.1007/s42979-024-03629-z

ORIGINAL RESEARCH

(Semi-)automatic Extraction of Urban Planning Rules in French for Better Management of Land Artificialization

Maksim Koptelov[1,2,3] · Margaux Holveck[4] · Bruno Cremilleux[1] · Justine Reynaud[1] · Mathieu Roche[3,5] · Maguelonne Teisseire[2,3]

[Koptelov *et al.* SN 2025]

Embeeding ciblé sur les termes avec élongation pour l'analyse de sentiment

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18786-9

Fusion of BERT embeddings and elongation-driven features

Abderrahim Rafae[1] · Mohammed Erritali[1] · Mathieu Roche[2,3]

Received: 18 September 2023 / Revised: 3 February 2024 / Accepted: 25 February 2024

[Rafae *et al.* MTA 2024]

Plan de la présentation

4. Quels sont les **nouveaux défis pluridisciplinaires** en particulier dans les pays du Sud ?

→ **Données :** Données peu dotées

→ **Méthode :** Production de nouveaux modèles

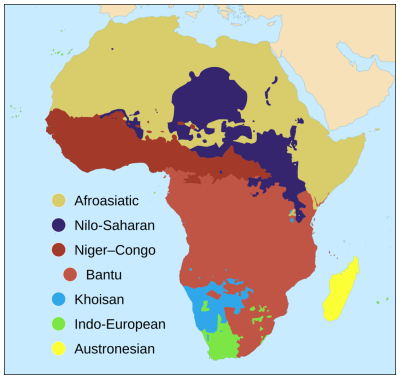
Plan de la présentation

4. Quels sont les **nouveaux défis pluridisciplinaires** en particulier dans les pays du Sud ?

→ **Données :** Données peu dotées

→ **Méthode :** Production de nouveaux modèles

NLP et Sud : Vers de nouveaux défis



Et les ressources peu dotées ?

AfricaNLP workshop at ICLR2023

ADAPTING TO THE LOW-RESOURCE DOUBLE-BIND: INVESTIGATING LOW-COMPUTE METHODS ON LOW-RESOURCE AFRICAN LANGUAGES

∀*, **Colin Leong**[1], **Herumb Shandilya**[*], **Bonaventure F. P. Dossou** [2,3,4,14], **Atnafu Lambebo Tonja**[5], **Joel Mathew**[6], **Abdul-Hakeem Omotayo** [7], **Oreen Yousuf**[*], **Zainab Akinjobi** [8], **Chris Chinenye Emezue** [9,14], **Shamsudeen Muhammad** [10], **Steven Kolawole** [11], **Younwoo Choi** [12], **Tosin Adewumi** [13]

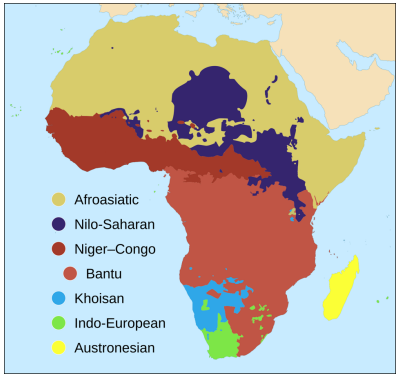
[*]Masakhane NLP, [1]University of Dayton, [2]Center for Intelligent Machines, McGill University, [3]Mila Quebec AI Institute, [4]Lelapa AI, [5]Instituto Politécnico Nacional, [6]USC Information Sciences Institute, [7]University of California, Davis [8]New Mexico State University, [9] Technical University of Munich, [10]University of Porto, [11]ML Collective, [12]University of Toronto, [13]ML Group, Luleå University of Technology,[14]Lanfrica.

Language (ISO)	Family	Region	Language Adapter Data		NER Finetuning Data		
			Train	Dev	Train	Dev	Test
Amharic (amh)	Afro-Asiatic-Ethio-Semitic	East	1037	899	1750	250	500
Fon (fon)	Niger-Congo-Volta-Niger	West	2637	1227	4343	621	1240
Hausa (hau)	Afro-Asiatic-Chadic	West	5865	1300	1903	2072	545
Igbo (ibo)	Niger-Congo-Volta-Niger	West	6998	1500	2233	319	638
Kinyarwanda (kin)	Niger-Congo-Bantu	East	1006	460	2110	301	604
Luganda (lug)	Niger-Congo-Bantu	East	4075	1500	2003	200	401
Nigerian-Pidgin (pcm)	English Creole	West	4790	1484	2100	300	600
Swahili (swa)	Niger-Congo-Bantu	East & Central	30782	1791	2104	300	602
Akan/Twi (twi)	Niger-Congo-Kwa	West	3337	1284	4240	605	1211
Wolof (wol)	Niger-Congo-Senegambia	West	3360	1506	1871	267	536
Yorùbá (yor)	Niger-Congo-Volta-Niger	West	6644	1544	2124	303	608
Zulu (zul)	Niger-Congo-Bantu	South	3500	1239	5848	836	1670

Table 1: Languages with ISO 639-2 Code. Language adapter training data was taken from the MAFAND dataset. NER fine-tuning and Evaluation data was taken from the MasakhaNER and MasakhaNER 2.0 datasets.

[Leong *et al.* AfricaNLP 2023]

NLP et Sud : Vers de nouveaux défis



Et les ressources peu dotées ?

Cheetah: Natural Language Generation for 517 African Languages

Ife Adebara[ξ,⋆] **AbdelRahim Elmadany**[ξ,⋆] **Muhammad Abdul-Mageed**[ξ,Ω,λ]
[ξ]The University of British Columbia [Ω]MBZUAI [λ] Invertible AI
{ife.adebara,a.elmadany,muhammad.mageed}@ubc.ca

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12798–12823
August 11-16, 2024

Category	LM	Lang/Total	African Languages	Families
Multilingual	MBART	3/50	afr, swh, yor.	2
	MT0	14/101	afr, amh, hau, ibo, lin, mlg, nyj, orm, sot, sna, som, swh, xho, yor, and zul	4
	MT5	12/101	afr, amh, nya, hau, ibo, mlg, sna, som, swh, xho, yor, and zul	3
African	AfriVeTa	10/10	gaz, amh, Gahuza, hau, ibo, pcm, som, swa, tir, and yor.	3
	AfriMT5	17/17	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul.	3
	AfriByT5	17/17	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul.	3
	AfriMBART	17/17	afr, amh, nya, hau, orm, som, swh, xho.	3
	Cheetah	517/517	Includes 517 African languages.	14

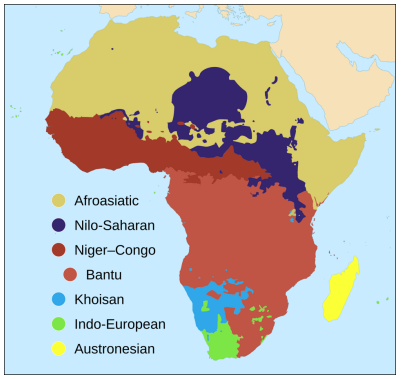
Table 1: Comparing with available encoder-decoder models with African languages represented. **Lang/Total**. describe the number of African languages comparing with the covered languages in the pretrained language models. **Families**. describes the number of covered language families.

AfroNLG, a multi-lingual, multi-task benchmark:

- Cloze Test
- Machine Translation
- Paraphrase
- Question Answering
- Summarization
- Title generation

[Adebara *et al.* EMNLP 2024]

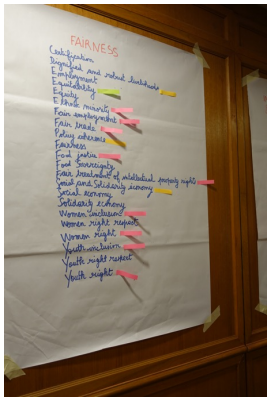
NLP et Sud : Vers de nouveaux défis

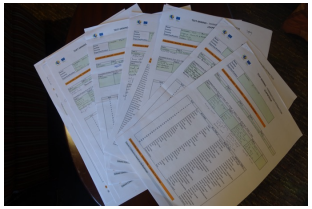


Quelques défis

- Données écrites sous-représentées
- Données hétérogènes
- Traitement de domaine de spécialité complexe avec de nombreux dialectes à considérer
- Biais de la collecte et de la sémantisation

ASSET





Plan de la présentation

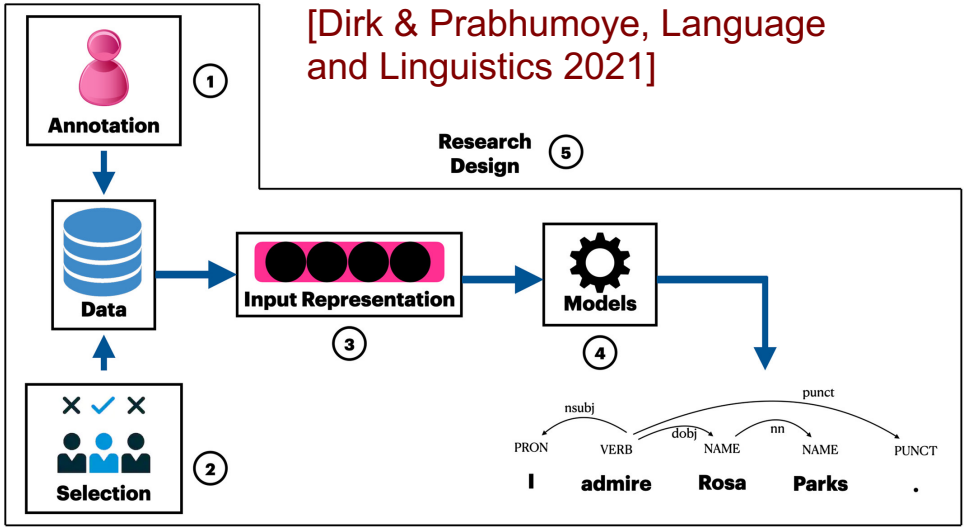
4. Quels sont les **nouveaux défis pluridisciplinaires** en particulier dans les pays du Sud ?

→ **Données :** Données peu dotées

→ **Méthode :** Production de nouveaux modèles

Production de nouveaux modèles

Et les biais ?



[Dirk & Prabhumoye, Language and Linguistics 2021]

- Biais liés aux méthodes d'augmentation

[Decoupes *et al.* Journal IDA 2024]

- Biais des modèles sur le volet spatial

[Decoupes *et al.* DS 2024]

Quelques défis

- Explicabilité
- Modèles et frugalité
- Détection de biais des modèles

Llama

400 milliards de paramètres

deepseek

600 milliards de paramètres

Merci ☺