

Des **TEXTES** et des réseaux

Nanée Chahinian

et

Equipe STARWARS : Mitra Aelami, Franco Albreto Cardillo, Serge Conrad, Franca Debole, Francesca Frontini, Batoul Haydar, Maria Keet, Ouafae Nahli, Ilham Chaker.

Equipe MeDo : Thierry Bonnabaud La Bruyère, Serge Conrad, Carole Delenne, Francesca Frontini, Julien Marin, Rachel Pankhurst, Mathieu Roche, Lucile Sautet, Maguelone Tesseire.

Contexte

2013 : Peu de données disponibles sur la structure des reseaux d'assainissement.

“Avec tout ce qui traîne sur le web, on pourrait trouver des infos sur les reseaux, non ?”

2015 : Projet Cart'eaux (Cartographie des réseaux d'Assainissement)

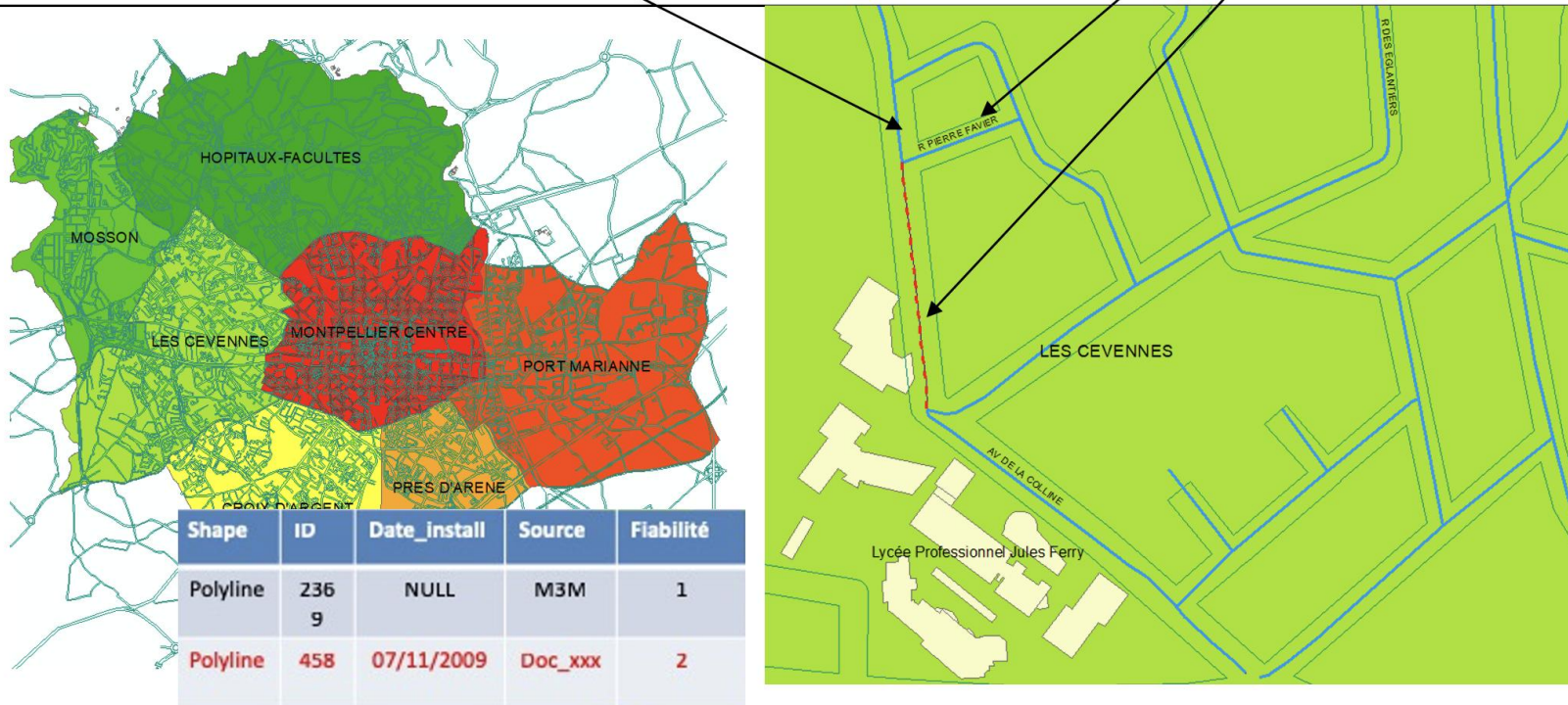
2018 : Projet MeDo (Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement).

2022 : Projet STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management)

Dans un monde idéal...

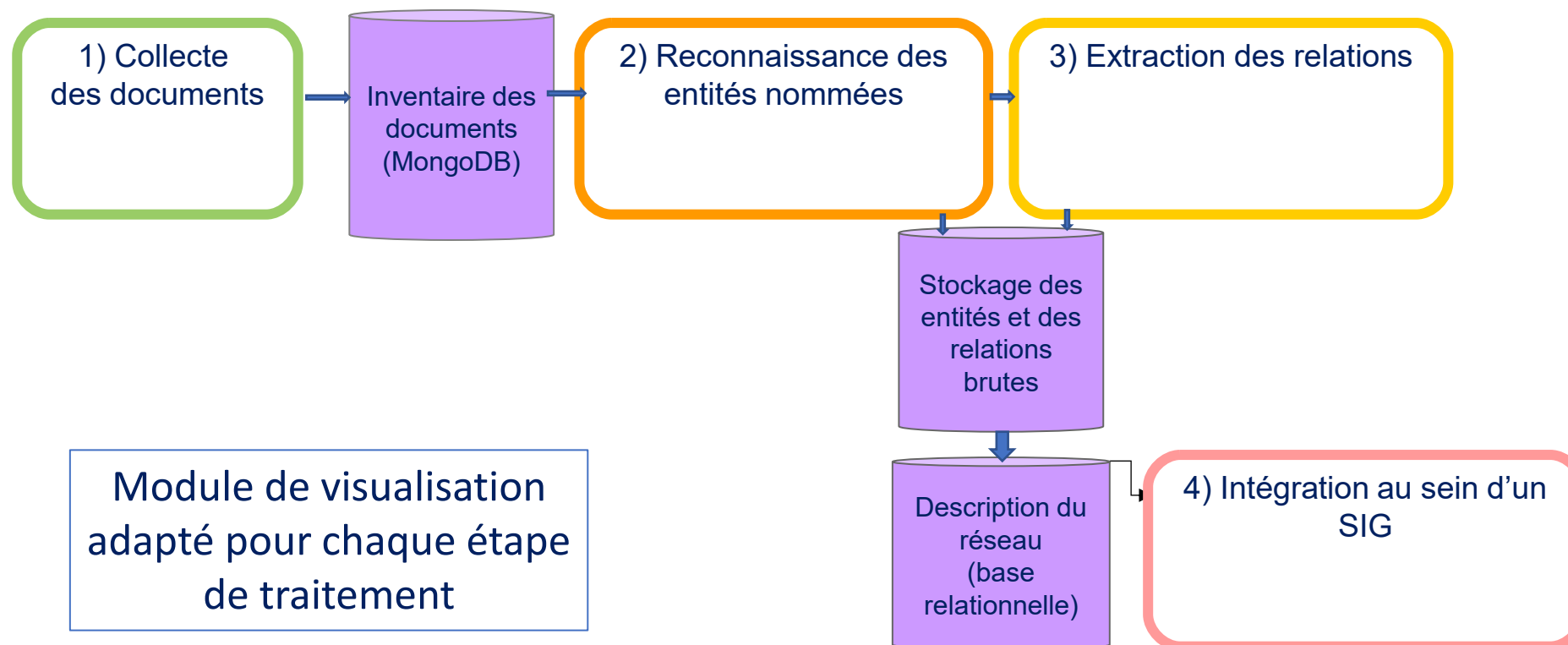
La **Métropole** va procéder à des travaux d'extension du **réseau d'eaux pluviales** **avenue de la Colline** dans le **quartier Cévennes**, à partir du **jeudi 20 octobre 2009**.

Ces **travaux**, qui débutent pour une **durée de 3 semaines**, consistent à réaliser une **extension** de la **conduite d'évacuation des eaux pluviales** de l'**avenue de la Colline**, notamment pour **raccorder** le **nouveau bâtiment** du lycée Jules Ferry. Cette **conduite** s'arrête **actuellement** au-dessus de la **rue Favier** et elle sera **prolongée** jusqu'à la **rue des Eglantiers**.



Bonnabaud La Bruyère et al., 2022

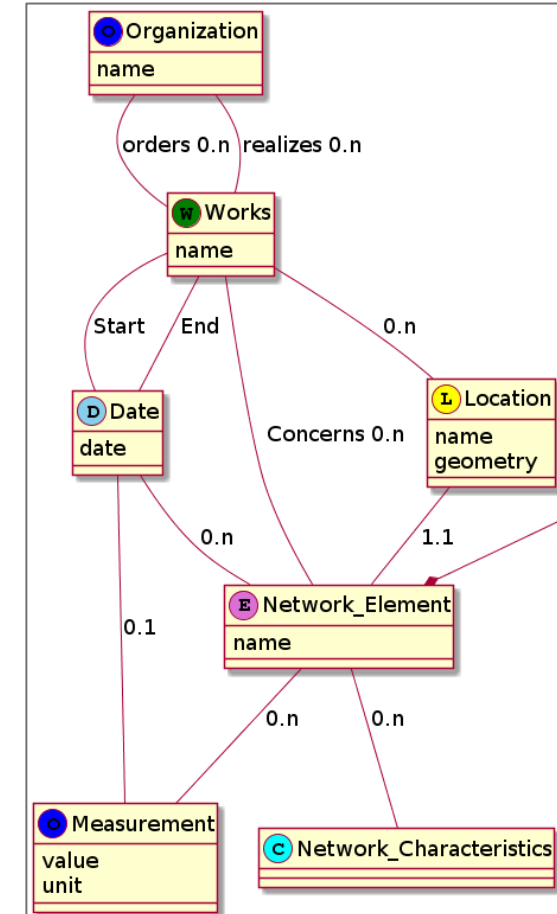
Plateforme WEIR-P



* Wastewater InfoRmation extraction Platform

Qu'est ce qu'on annote ?

- Entités nommées spécifiques au domaine
 - Réseau : type, élément, caractéristiques.
 - Événements: Accidents & travaux
 - Temporel et spatial
 - Organisations
 - Mesures
 - Station d'épuration
 - Qualité d'eau
 - Gestion
 - Indicateur
- 23 documents annotés manuellement sous BRAT i.e. 1,387 phrases et 4,505 entités (<https://brat.nlplab.org/>).
- Gold standard et corpus disponible
https://dataverse.ird.fr/dataverse/project_medo



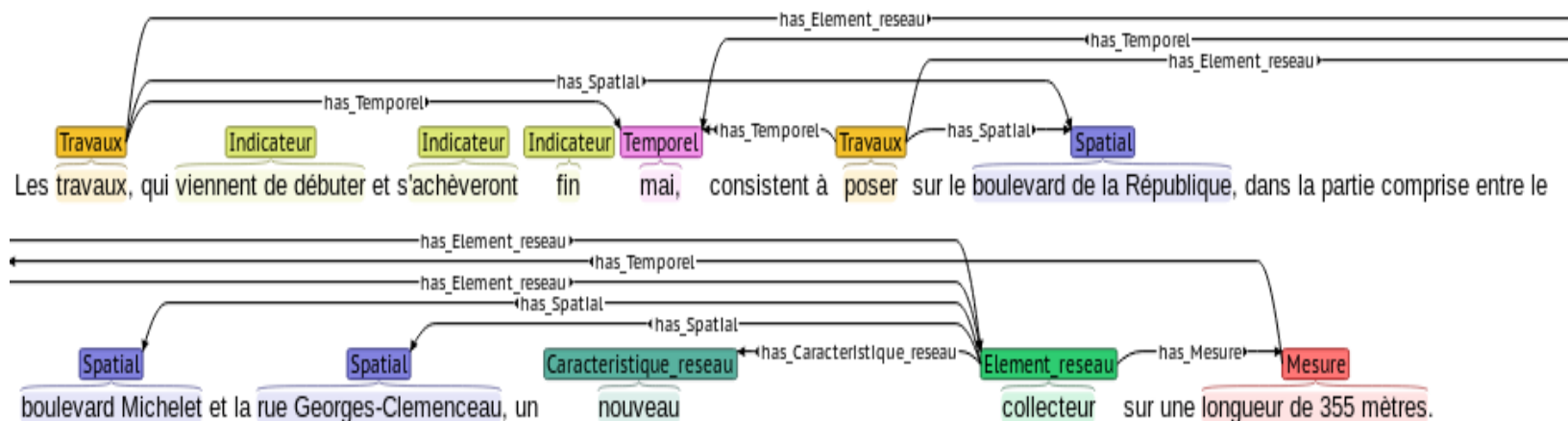
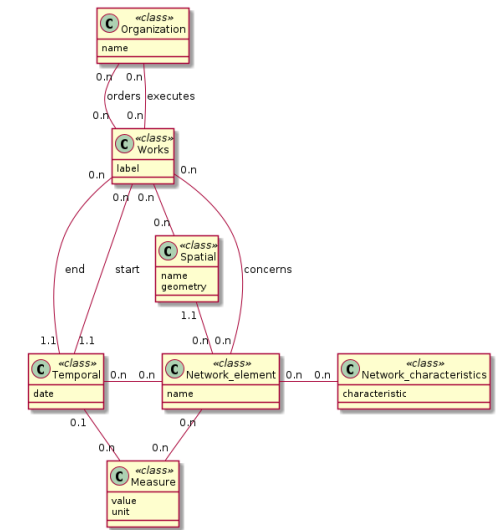
Méthodologie : Extraction d'entités nommées

- Classification de relevance : Scikit-learn, Multinomial Naive Bayes, IDF-TF
- Découpage des phrases, tokenisation et étiquetage morpho-syntaxique : TreeTagger
- Évaluation : MUC-5 (Chinchor& Sundheim, 1993)
- Ensemble module REN : 13 entités
 - Core-NLP (meilleure precision, F-score= 50.42) ⇒ En cas de conflit, choix du modèle le plus performant
 - SpaCy (Meilleur rappel, F-score=52.29) ⇒ Pour entités temporelles, Heideltime
 - Heideltime (temporels) outrepasse les autres.

Adaptation d'outils existants au domaine de l'assainissement urbain

Méthodologie : Extraction des relations

- Sur la base d'un modèle conceptuel développé par le projet MeDo.
- Extraction des relations depuis chaque entité avec :
 - Motifs syntaxiques
 - Annotations environnantes (jusqu'à 2 phrases selon type)
 - Titre du document
- Liage spatial : Base Adresse Nationale, Geonames & Nominatim

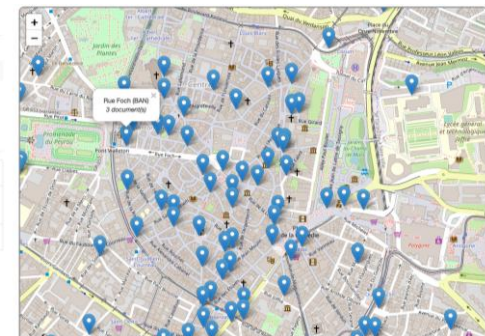


Sélection

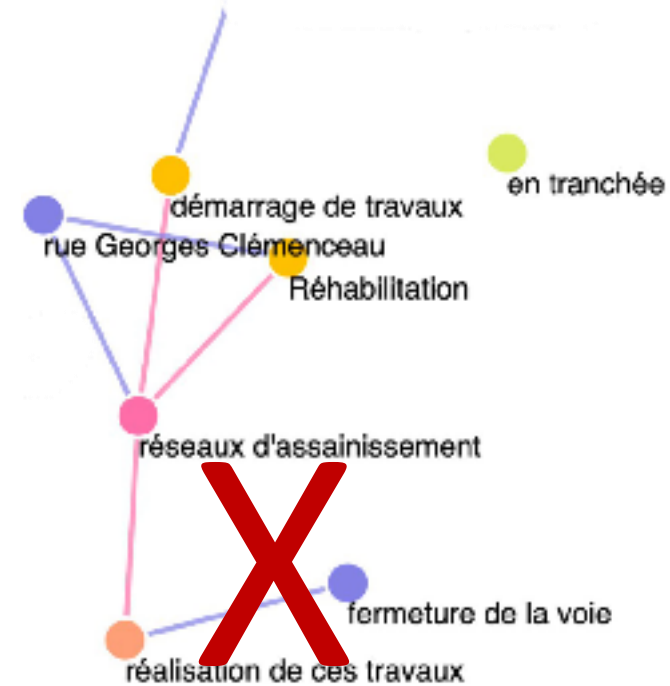
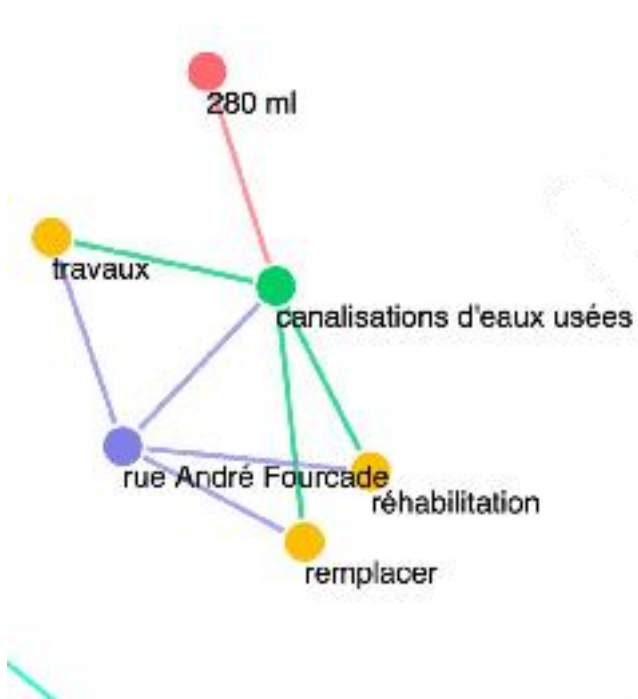
Sélectionnez un corpus
web_montpellier_19092019

Les documents associés à :
Rue Foch (BAN)

Transports_en_commun_de_Montpellier
modernisation_des_reseaux_d_eau_potable_et_assainissement_a_murviel_1510242.php
e98b4326_3120_d8aa_8313_a0c3d7b78ee2



Résultats : Exemple de mise en relation



Temporel

Spatial

Organisation

Indicateur

Travaux

Traitement

Accident

Type_reseau

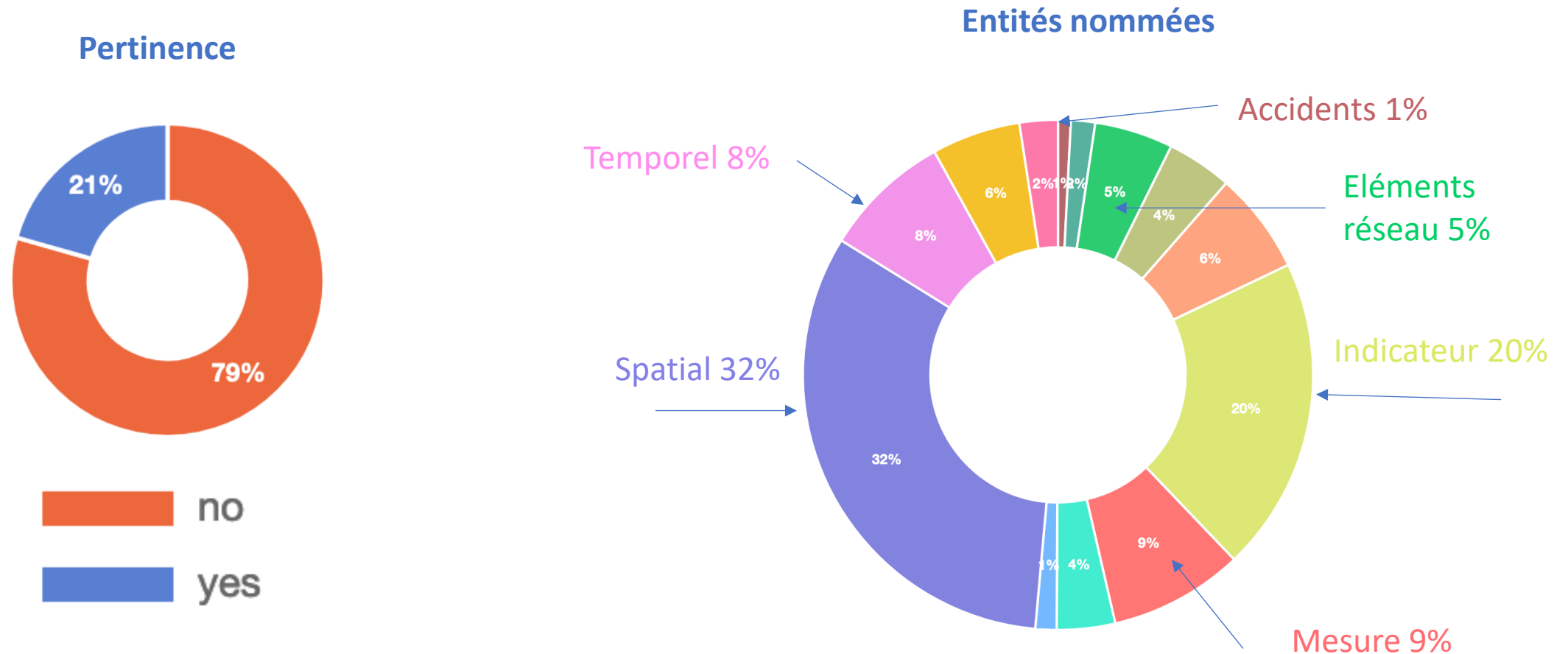
Element_reseau

Gestion

Mesure

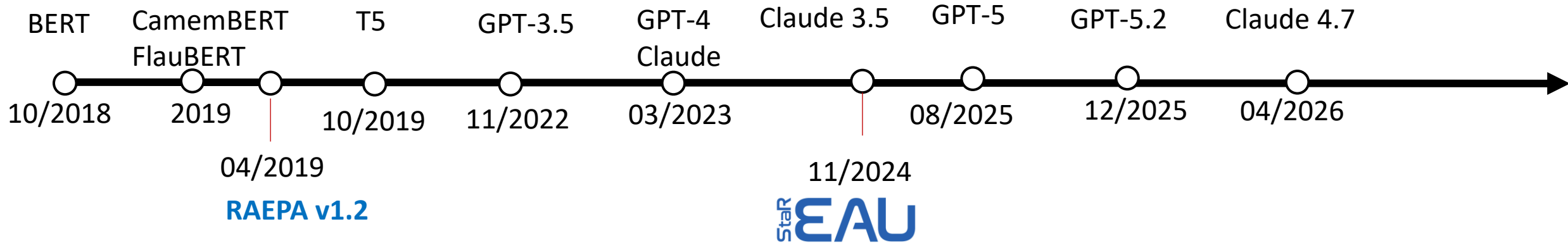
Validation : JACOU

Jacou (Hérault), collecté en juillet 2020, 728 documents, 59,318 EN, 42359 relations

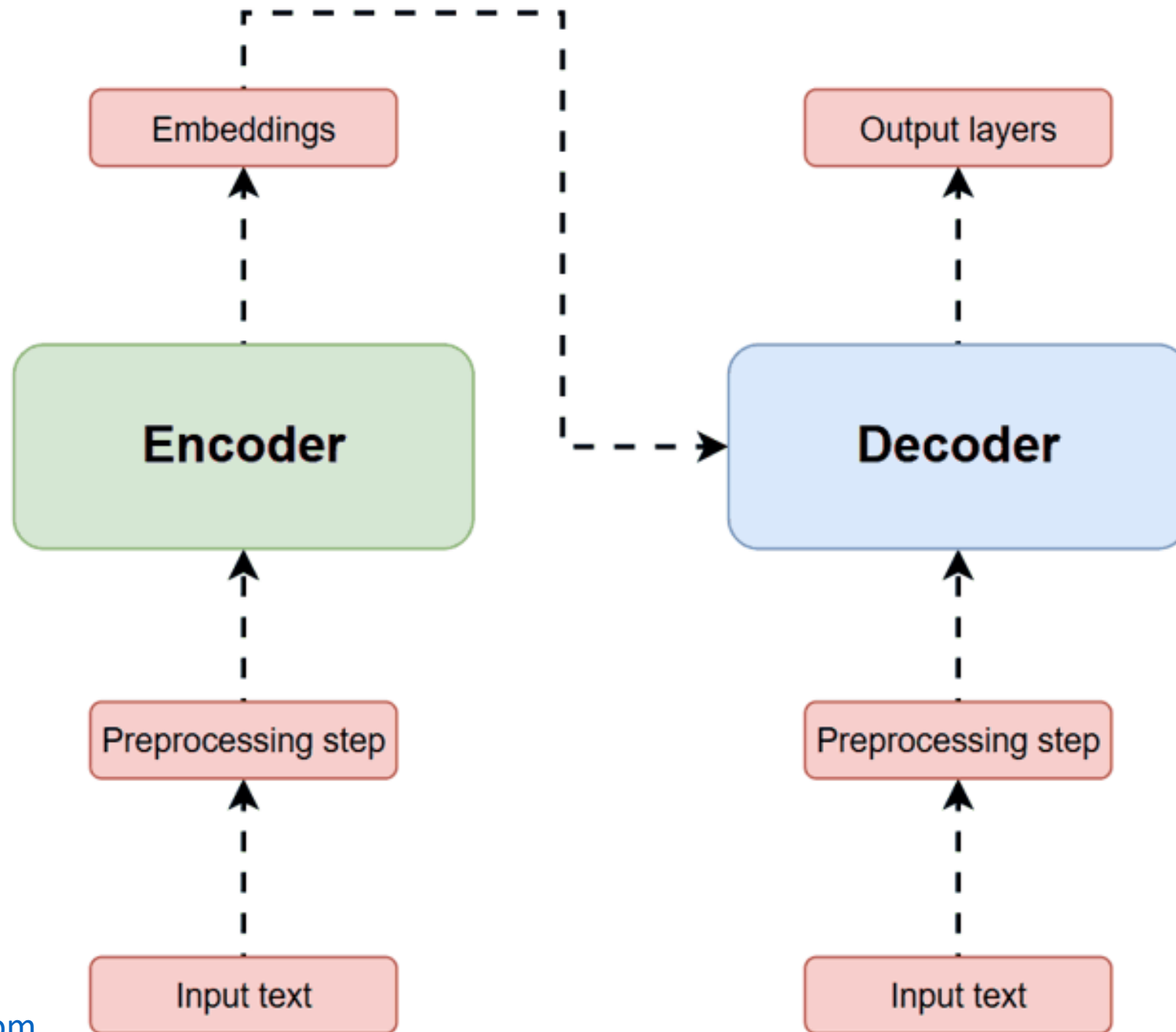


Résultats variables selon type entités. Nécessité d'améliorer le liage spatial

Pendant ce temps...



- Apparition des transformers : les traitements se font en parallèle => plus rapides.
- Introduction d'encodeurs
 - Mécanismes d'auto-attention => le modèle se concentre sur les mots les plus pertinents.
 - Réseau neuronal feedforward => le modèle saisit des relations plus profondes entre les mots.
- T5 (Text-to-Text Transfer Transformers) pré-entraînés sur le Colossal Clean Crawled Corpus moissonné du web



De l'approche monolingue au multilingue avec des LLMs



IN

Travaux précédents

- **MeDo** Corpus Français
- **MeDo** guide d'annotation

Corpus Français

- **1,557 documents**
HTML/PDF documents collected
- **Gold standard de 80 documents**

OUT

Ressources

Corpus de textes
Français/Italien/Anglais

*110 documents, dont 45 du
Gold Standard MeDo*

Benchmark

Corpus aligné Français,
Italien, Anglais.

Expérience de REN
multilingue

Arabe

Corpus de textes
réglementaires.

2 lois et décrets associés

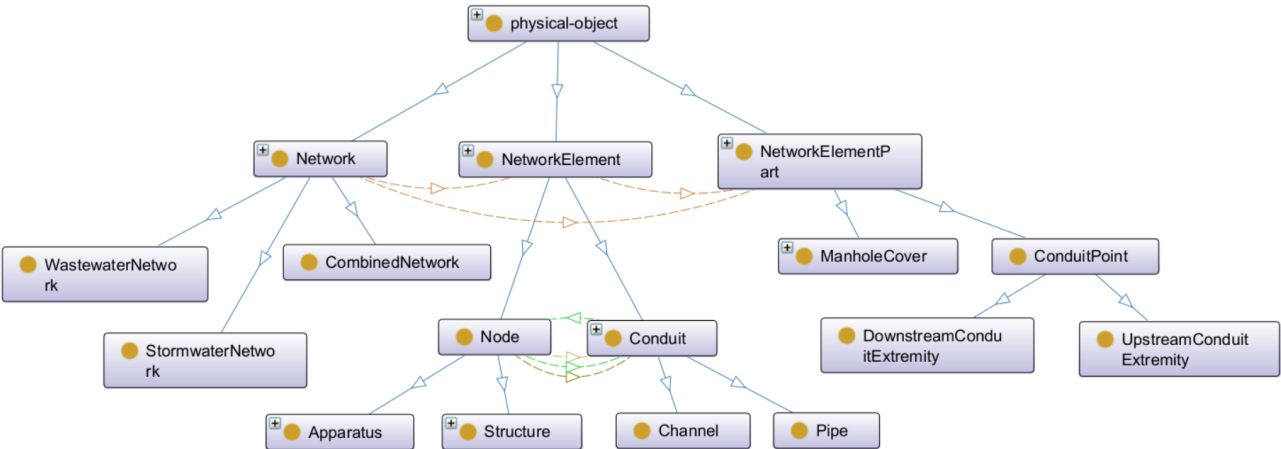
Plateforme bilingue - lecteur

[https://yangkai.pyt
honanywhere.com/](https://yangkai.pythonanywhere.com/)



Nouveau modèle conceptuel de données et ontologie

Types d'entités (MeDo 2018)	Nos types d'entités (2024)
Type_reseau	Type_reseau
Organisation	Organisation
Spatial	Spatial
Mesure	Mesure
Temporel	Temporel
Epuration	Canalisation
Element_reseau	Ouvrage
Travaux	Appareillage
Qualite_eau	Matériau
Caracteristique_reseau	Forme
Gestion_reseau	Mode_circulation
Indicateur	Evénements
Accident	Fonction
	Defaillance



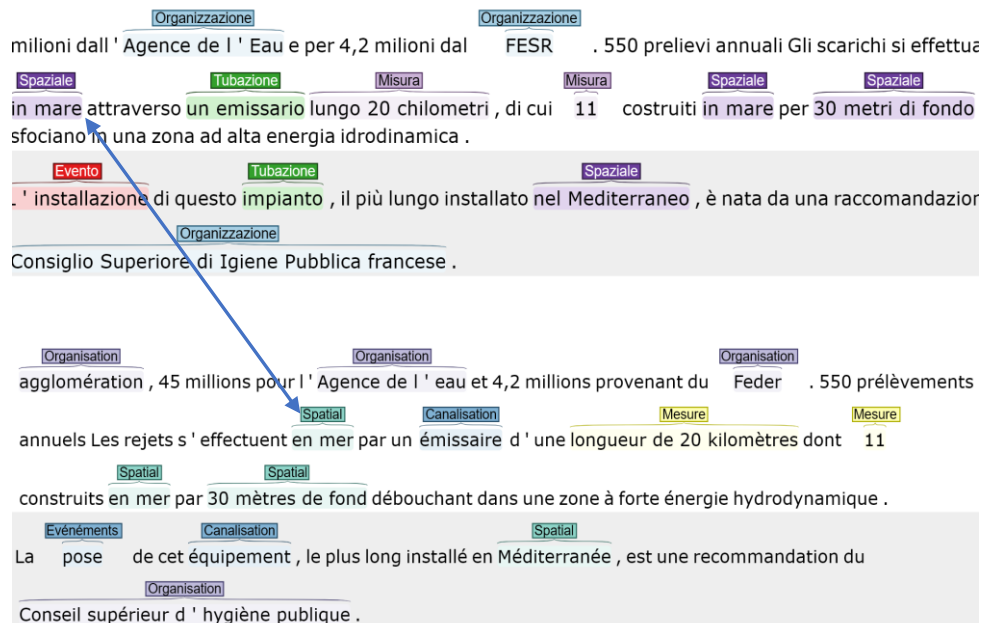
110 documents annotés dont 45 du Gold Standard MeDo

Des méthodes différentes selon les langues

• Italien

Traduction des documents et des catégories d'annotation

- Etape 1 : automatique
- Etape 2: étudiants bilingues et hydrologue



Debole, 2026

• Arabe

Correspondance lexicale en usage dans le domaine

- Constitution d'un corpus bilingue : 2 texte de loi+decrets associés
- Encodage structuré en TEI XML
- Analyse des usages en contexte
- Identification des correspondances terminologiques FR–AR
- Annotation terminologique <term>, @ana
- Extraction des correspondances terminologiques FR–AR
- Validation par experts du domaine

```
<item xml:id="fr_art3_7" corresp="ar_art3_7">
  <p>
    <cit type="parallel">
      <quote>
        <term xml:id="fr_eau_usee_def"
          corresp="#ar_eau_usee_def" ana="#def">eau usée</term> : une eau qui a
          subi une modification de sa composition ou de son état naturel du fait
          de son utilisation ; </quote>
      <quote>
        <term xml:id="ar_eau_usee_def"
          corresp="#fr_eau_usee_def" ana="#def">ماء مستعمل</term>: ماء تعرض لتغيير
          في تركيبته أو حالته الطبيعية جراء استعماله؛ </quote>
      </cit>
    </p>
  </item>
```

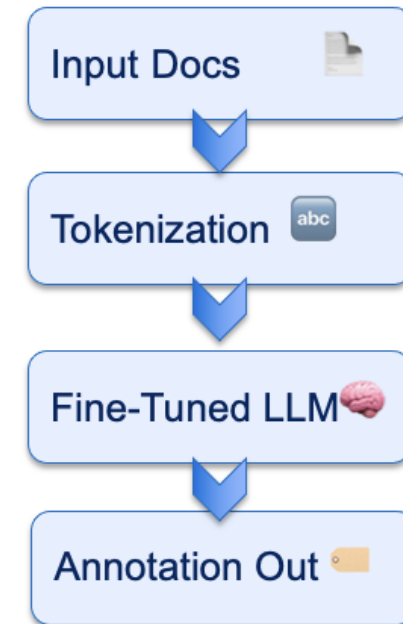
Nahli et al., 2026

Extraction d'entités nommées

- Les LLMS

- CamemBERT (110M param -- 138GB -- 32.7B tokens)
- GiBERTo (110M param -- 71GB -- 11B tokens)
- ItalianBERT (110M param --13GB -- 2B tokens)

- 70%–10%– 20% entraînement/validation/test.
- Set Français et Italien contiennent les mêmes documents.
- Suite strictement croissante de sous-ensembles d'entraînement
- Pour chaque taille d'échantillon d'entraînement, plusieurs runs. L'ensemble test reste identique
- 40 epochs



Gold Standard



LLM vs Gold Labels



F1 0.63



CamemBERT

F1 0.53



ItalianBERT

F1 0.50



GiBERTo

Besoin de plus de données !

Utilisation de prompts pour projeter des annotations



System Prompt

«You are an expert in the wastewater and stormwater domain and in named entity recognition»

Modèles

- Llama2,
- Llama3 (Meta);
- Mistral (Mistral.ai);
- Phi4 (Microsoft).

User Prompt

«Your task is to transfer IOB2 tags from a source sentence in [SRC] to its aligned [TRG] translation. Only transfer the tags; do not invent new entities.»

«Output only the annotated target sentence. Do not explain your answer. Do not include reasoning, commentary, or formatting.»

- The source sentence is: [source sentence]
- The IOB2 tags for the source are: [list of tags]
- The aligned [TRG] sentence is: [target sentence]
- Return the [TRG] sentence with transferred IOB2 tags in the same format.
- Respond only with the tagged sentence: [[first target token, "..."], . . .]

Résultats 1shot – 1 instruction avec 1 exemple

	Model	Format tag	Tags FR → IT	F1 FR → IT	Tags IT → FR	F1 IT → FR	Observation
→	Gemma3 27.4B	HTML ±	14	0.57	14	0.58	Correct tags
	Gemma3 27.4B	IOB2	15	0.62	Fails	Fails	it fails the task *
	Llama3 70.6B	HTML	38	0.58	24	0.64	“invent” new tags *
	Llama3 70.6B	IOB2	14	0.67	15	0.62	“invent” new tags *
→	Mistral 122.6B	HTML ±	14	0.51	14	0.53	Correct tags
	Mistral 122.6B	IOB2	16	0.64	15	0.57	“invent” new tags *
	Phi4	HTML	16	0.53	16	0.52	“invent” new tags *
	Phi4	IOB2	26	0.58	27	0.41	“invent” new tags *

Certains modèles inventent de nouvelles catégories de tags, malgré l'instruction
Le format des tags influe les résultats. HTML semble plus facile à contrôler que IOB2

Même les modèles larges buttent face à des instructions inhabituelles

Conclusions et perspectives

- Malgré les avancées technologiques, la fouille de données textuelle reste compliquée
- Toujours besoin de grands échantillons or le sujet n'emballe pas les foules
- Peu d'informations quantitatives, plus utile pour traquer des événements que déterminer la géométrie
- Les LLMs hallucinent et c'est normal



Références

- Franco Alberto Cardillo, Franca Debole, Francesca Frontini, Mitra Aelami, Nanée Chahinian, et al.. Novel Benchmark for NER in the Wastewater and Stormwater Domain. *8th IEEE CiSt'25 Congress*, Oct 2025, Marrakech (Maroc), Morocco. pp.226-231, [10.1109/CiSt65886.2025.11224095](https://doi.org/10.1109/CiSt65886.2025.11224095). [hal-05102736](https://hal.archives-ouvertes.fr/hal-05102736)
- Nanée Chahinian, Thierry Bonnabaud La Bruyère, Francesca Frontini, Carole Delenne, Marin Julien, et al.. WEIR-P: An Information Extraction Pipeline for the Wastewater Domain. *RCIS 2021 - 5th International Conference on Research Challenges in Information Science*, May 2021, Online, Cyprus. [10.1007/978-3-030-75018-3_11](https://doi.org/10.1007/978-3-030-75018-3_11). [hal-03211461](https://hal.archives-ouvertes.fr/hal-03211461)
- Zola Mahlaza, C Maria Keet, Nanée Chahinian, Batoul Haydar. On the Feasibility of LLM-based Automated Generation and Filtering of Competency Questions for Ontologies. *LDK'25, Proceedings of the 5th Conference on Language, Data and Knowledge*, Sep 2025, Naples, Italy. pp.136. [hal-05284209](https://hal.archives-ouvertes.fr/hal-05284209)
- Ouafae Nahli, Nanee Chahinian, Ilham Chaker. Corpus et terminologie bilingue arabe-français du domaine eaux usées. *TOTH 2026 Terminology & Ontology: Theories and applications*, Université de Crète – TALOS AI for SSH Numéro de Projet 101087269, Jun 2026, Chamberry, France. [hal-05655822](https://hal.archives-ouvertes.fr/hal-05655822)